

Free Energy Perturbation Hamiltonian Replica-Exchange Molecular Dynamics (FEP/H-REMD) for Absolute Ligand Binding Free Energy Calculations

Wei Jiang[†] and Benoît Roux^{*,†,‡}

Biosciences Division, Argonne National Laboratory, 9700 South Cass Avenue, Building 240, Argonne, Illinois 60439, and Department of Biochemistry and Molecular Biology, Gordon Center for Integrative Science, University of Chicago, 929 57th Street, Chicago, Illinois 60637

Received April 1, 2010

Abstract: Free Energy Perturbation with Replica Exchange Molecular Dynamics (FEP/REMD) offers a powerful strategy to improve the convergence of free energy computations. In particular, it has been shown previously that a FEP/REMD scheme allowing random moves within an extended replica ensemble of thermodynamic coupling parameters “ λ ” can improve the statistical convergence in calculations of absolute binding free energy of ligands to proteins [*J. Chem. Theory Comput.* 2009, 5, 2583]. In the present study, FEP/REMD is extended and combined with an accelerated MD simulations method based on Hamiltonian replica-exchange MD (H-REMD) to overcome the additional problems arising from the existence of kinetically trapped conformations within the protein receptor. In the combined strategy, each system with a given thermodynamic coupling factor λ in the extended ensemble is further coupled with a set of replicas evolving on a biased energy surface with boosting potentials used to accelerate the interconversion among different rotameric states of the side chains in the neighborhood of the binding site. Exchanges are allowed to occur alternatively along the axes corresponding to the thermodynamic coupling parameter λ and the boosting potential, in an extended dual array of coupled λ - and H-REMD simulations. The method is implemented on the basis of new extensions to the REPDSTR module of the biomolecular simulation program CHARMM. As an illustrative example, the absolute binding free energy of *p*-xylene to the nonpolar cavity of the L99A mutant of the T4 lysozyme was

calculated. The tests demonstrate that the dual λ -REMD and H-REMD simulation scheme greatly accelerates the configurational sampling of the rotameric states of the side chains around the binding pocket, thereby improving the convergence of the FEP computations.

Introduction

Free energy perturbation molecular dynamics (FEP/MD) simulations with explicit solvent molecules provide one of the most fundamental routes for computing the binding affinities of small compounds to proteins.^{1,2} In practice, a critical issue with FEP/MD simulations is to achieve a sufficient sampling of all the relevant degrees of freedom. Problems can arise with large structural reorganizations either in the ligand or in the protein upon formation of the bound complex because sampling those is typically beyond the reach of straight brute-force FEP/MD simulations. More specifically, when there are large energy barriers separating the relevant conformational states, the ligand or the protein may remain kinetically trapped in the starting configuration for a very long time during FEP/MD simulations, and alternate conformations are never visited. The incomplete configurational sampling results in computed binding free energies that are dependent on the starting protein or ligand configuration, which are of limited significance and practical use.

The structural changes observed upon the binding of aromatic molecules to a nonpolar cavity engineered in the L99A mutant of the T4 lysozyme (T4L) provide a good illustration of the type of problems that can arise from insufficient sampling (Figure 1). For the bound complexes involving small and medium-sized ligands (e.g., benzene, toluene, benzofurane, and indole), the protein structure is essentially identical to the ligand-free (*apo*) conformation. For those ligands, the calculated absolute binding free energies are well converged, regardless of whether the FEP/MD simulations are started from the *holo* or the *apo* state.^{1,3,4} Difficulties arise in the case of larger ligands (e.g., indene, *n*-butylbenzene, isobutylbenzene, *o*-xylene, and *p*-xylene). In this case, the side chain of Val111, which is in direct contact with the bound ligand, changes its rotameric states from a *trans* conformation ($\chi_1 = 180^\circ$) for the ligand-free *apo* to a *gauche* conformation ($\chi_1 = -60^\circ$) for the bound state with large ligands. The intrinsic energy barrier around the χ_1 torsion of the valine (~ 5 kcal/mol) is sufficient to prevent the side chain from reorienting on the time scale of typical FEP/MD simula-

* Corresponding author e-mail: roux@uchicago.edu.

[†] Argonne National Laboratory.

[‡] University of Chicago.

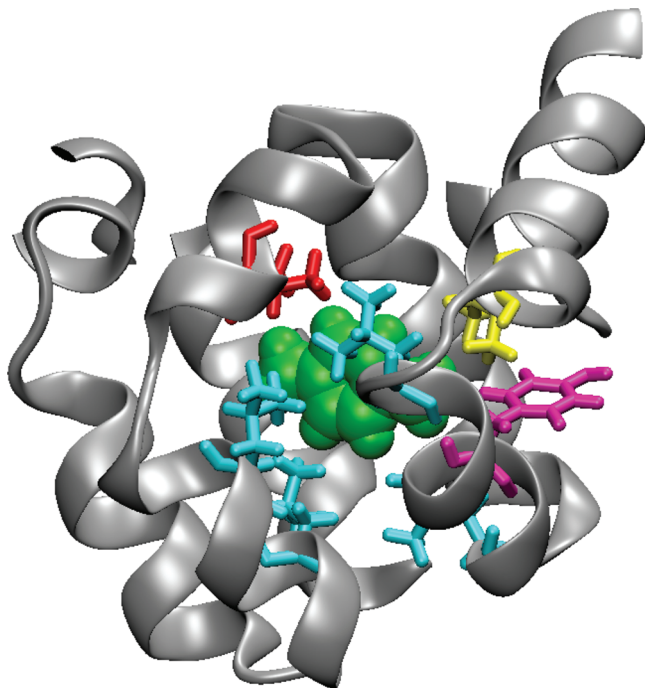


Figure 1. The artificially engineered nonpolar cavity of the L99A mutant of the T4 lysozyme (T4L/L99A) with *p*-xylene bound. Highlighted are seven protein side chains within 6 Å of the ligand (PDB 187L). Red color: valine 111. Blue color: leucine 84, 91, 118, and 121. Purple color: tyrosine 88. Yellow color: isoleucine 78.

tions. As a consequence, a FEP/MD calculation started from the *holo* state with the Val111 in the *gauche* state remains kinetically trapped while the ligand is alchemically decoupled, yielding a calculated binding free energy that is too favorable by 2–3 kcal/mol.¹ Alternatively, a FEP/MD calculation started from the *apo* state is too unfavorable by 2 kcal/mol.⁴ As discussed in detail by Mobley et al.,⁴ the lack of consistency between the two series of FEP calculations directly reflects the incomplete configurational averaging caused by the slow relaxation of the kinetically trapped degrees of freedom.

An elegant and powerful approach to enhance the sampling of the slowly varying degrees of freedom is to introduce a restraining potential serving as a “guide” to help reduce the size of the configurational space that needs to be explored during the free energy calculation. In practice, this first requires the identification of a key order parameter, ξ , associated with the slowly varying structural feature. Then, the potential of mean force (PMF) along this order parameter, $W(\xi)$, must be calculated via umbrella sampling biased simulations, and standard alchemical FEP/MD calculations are carried out in the presence of a biasing potential restricting the dynamics along ξ over a small range. Finally, unbiased thermodynamic averages for the entire association/dissociation process can be obtained by carrying out the explicit numerical integration of the probability distributions involving the Boltzmann factor of the PMF, $\exp[-W(\xi)/k_B T]$.^{1,2,4–6} The free energy difference is evaluated as the reversible work for switching on a conformational restraint in one end-point state and switching it off in the other, according to a so-called “confine-and-release” cycle.⁴ One might refer to this entire procedure as a “deliberate” PMF-based sampling strategy.

One important drawback from a deliberate PMF-based sampling strategy is that it relies on the prior identification of one or a few key degrees of freedom that one intends to control via umbrella sampling simulations. In the general case, it may not always be easy to determine which degrees of freedom might be slowly varying. A possible route to resolve the situation could be to extend the PMF-based strategy to multiple order parameters, but in practice, this does require carrying out umbrella sampling simulations for the entire multidimensional subspace. Thus, a deliberate PMF-based sampling strategy becomes rapidly unwieldy and inapplicable in the general situation where there could be structural rearrangements involving many elements. An alternative approach might be to simulate a conveniently chosen artificial reference state with soft cores as in the enhanced sampling—one-step perturbation method (ES-OS),⁷ although it is unclear if this could treat multiple side chains surrounding a protein binding site simultaneously. A general treatment of structural relaxation remains a major challenge for applications in computer-assisted drug lead discovery,⁸ where the main source of structural information may be ligand docking.

An alternative to a deliberate PMF-based sampling strategy is to exploit the concept of accelerated MD to increase the interconversion rates between metastable states.^{9–12} The central element of accelerated MD consists in introducing a “boosting” potential that biases the energy surface to cancel out the intrinsic energy barriers opposing the relevant transitions that one wishes to sample. To retain the proper thermodynamic Boltzmann sampling of the system, the accelerated simulation can be combined with a parallel tempering Hamiltonian-REMD (H-REMD) scheme.^{13,14} While this approach also requires the prior identification of the relevant subspace corresponding to the slowly varying degrees of freedom, the method is considerably less computationally expensive than the need to perform umbrella sampling simulations over multiple degrees of freedom as with a PMF-based strategy. An adequate sampling of the relevant subspace is expected to be, in most case, computationally affordable via a H-REMD scheme. In particular, as exemplified by the isomerization of Val111 in the L99A mutant of T4L discussed above, transitions of side chains and/or backbones in the neighborhood of the binding pocket clearly dominate the structural relaxation of the protein receptor in ligand binding free energy computations. More generally, the total number of side chains in the neighborhood of a binding pocket is fairly limited and it is likely that their dynamical transitions could be accelerated with FEP/H-REMD simulations.

In a previous communication, free energy perturbation (FEP) with a staged reversible thermodynamic work protocol designed for the calculation of absolute ligand binding affinities was combined with a distributed replica exchange MD (λ -REMD) simulation scheme.¹⁵ It was shown that this FEP/ λ -REMD scheme could improve the statistical convergence of FEP calculations by allowing random Monte Carlo moves in an extended ensemble of thermodynamic coupling parameter λ . The important concept of replica exchange in binding free energy calculations was first introduced by Woods and co-workers.¹⁶ However, a straightforward FEP/ λ -REMD algorithm is insufficient to accelerate the sampling of kinetically trapped degrees of freedom such as the isomerization of Val111 in the L99A mutant of T4L. The exchanges along the thermodynamic

coupling λ can help to mix the side chain rotamers of the protein in the *apo* and *holo* states, but transitions occur rarely due to the intrinsic dihedral energy barriers. In the present work, we extend those ideas to propose a rapid and robust framework for free energy computations combining the concept of λ -REMD simulations within the staged FEP, and the concept of accelerated MD with boosting potentials via H-REMD. To achieve the proper sampling enhancement in the relevant subspace, we combine λ -REMD with H-REMD. Random moves are allowed within an extended set of replicas biased by different values of the boosting factor “ b ” controlling the amplitude of a biasing potential according to a H-REMD scheme. The implementation is based on the MPI level parallel/parallel mode made possible by the Distributed Replica (REPDSTR) technique^{17,18} of the program CHARMM,¹⁹ in which each λ window of FEP is treated as an independent replica with its private I/O. With REPDSTR, it is straightforward to introduce a set of auxiliary boosting replicas for each λ window. This yields a dual REMD protocol for FEP calculations, with replica-exchange along two axes (2D) corresponding to the thermodynamic coupling parameter λ and a second axis corresponding to the boosting factor b . The entire array of REMD simulations can be executed as a single job via REPDSTR. It is shown that the dual FEP simulation scheme combining λ -REMD and H-REMD significantly accelerates the configurational sampling of the protein in FEP calculations. The method is illustrated with the calculation of the absolute binding free energy of *p*-xylene to the nonpolar cavity of T4L/L99A.

Computational Details

A. REPDSTR Implementation of Staging Simulation Protocol. In the FEP staging simulation protocol, the potential energy is expressed in terms of four coupling (window) parameters^{1,2,20}

$$U(\lambda_{\text{rep}}, \lambda_{\text{dis}}, \lambda_{\text{elec}}, \lambda_{\text{rstr}}) = U_0 + U_{\text{rep}}(\lambda_{\text{rep}}) + \lambda_{\text{dis}} U_{\text{dis}} + \lambda_{\text{elec}} U_{\text{elec}} + \lambda_{\text{rstr}} U_{\text{rstr}} \quad (1)$$

where U_0 is the potential of the system with the noninteracting (decoupled) ligand; λ_{rep} , λ_{dis} , λ_{elec} , and $\lambda_{\text{rstr}} \in [0,1]$ are the thermodynamic coupling parameters; U_{rep} and U_{dis} are the shifted Weeks–Chandler–Anderson²¹ (WCA) repulsive and dispersive components of the Lennard-Jones potential; U_{elec} is the electrostatic contribution; and U_{rstr} is the restraining potential.

With the updated REPDSTR module of CHARMM,¹⁹ the four-stage FEP simulation protocol can be implemented in a single parallel/parallel MPI job. Figure 2a shows the REPDSTR implementation of the updated FEP/REMD scheme, which is able to support the complete insertion process of the ligand into the binding pocket. The free energy corresponding to the process of inserting the ligand into the binding site is

$$U(\lambda_{\text{rep}} = 0, \lambda_{\text{dis}} = 0, \lambda_{\text{elec}} = 0, \lambda_{\text{rstr}} = 1) \rightarrow U(\lambda_{\text{rep}} = 1, \lambda_{\text{dis}} = 1, \lambda_{\text{elec}} = 1, \lambda_{\text{rstr}} = 0) \quad (1a)$$

To achieve a significant sampling enhancement, M additional replicas with “boosting” biasing potentials are introduced ($b = 0, 1/M, 2/M, \dots, 1$) for each λ value of the FEP/REMD

calculation. The boosting parameter b scales the biasing potential (the system is not biased when $b = 0$, and the biasing potential cancels out the intrinsic dihedral PMF of the side chain when $b = 1$). Figure 2b shows the FEP/REMD scheme with a biased Hamiltonian. Replica exchanges are attempted alternatively in λ space and b space, forming a two-dimensional (2D) REMD framework.

The replica-exchange algorithm follows the conventional Metropolis Monte Carlo exchange criterion:

$$P(\lambda_i \rightarrow \lambda_j) = \min\{1, e^{(-[U(\lambda_i, b_0, \mathbf{r}_{i,0}) + U(\lambda_j, b_0, \mathbf{r}_{j,0}) - U(\lambda_i, b_0, \mathbf{r}_{i,0}) - U(\lambda_j, b_0, \mathbf{r}_{j,0})]) / k_B T}\} \quad (2)$$

$$P(b_k \rightarrow b_l) = \min\{1, e^{-[U(\lambda_i, b_k, \mathbf{r}_{i,k}) + U(\lambda_j, b_l, \mathbf{r}_{j,l}) - U(\lambda_i, b_k, \mathbf{r}_{i,k}) - U(\lambda_j, b_l, \mathbf{r}_{j,l})] / k_B T}\} \quad (3)$$

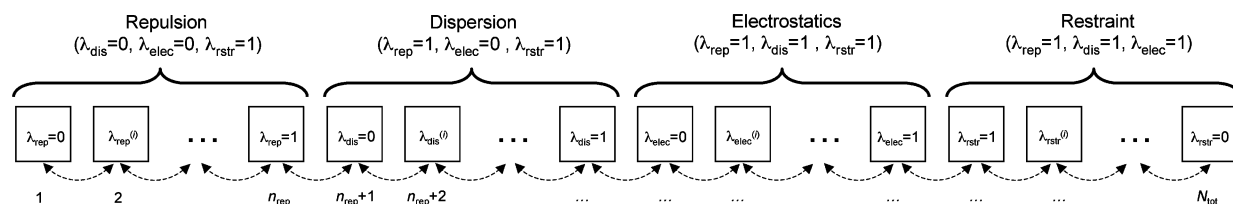
where U denotes the potential energy of the underlying replica, λ_i and λ_j denote the staging parameters, and b_k and b_l denote the boosting parameters.

B. Biasing Potentials for Residue χ_1 Dihedral Angle. Effective biasing boosting potentials can be obtained by fitting the potential of mean force (PMF) of individual dihedral degrees of freedom calculated on small peptides, in the spirit of the work of Kannan and Zacharis.¹⁰ In the present application, biasing potentials for the side chain dihedral angles χ_1 were constructed by calculating a PMF in the gas phase. The biasing potentials were determined for valine, leucine, isoleucine, and tyrosine residues using umbrella sampling simulations of one-residue peptides. In the ligand binding calculation, these types of residues are distributed within a distance of 6 Å away from the center of mass of the ligand. A series of quadratic umbrella potentials with a force constant of 100 kcal mol⁻¹ rad⁻² and distributed every 5° was used. The angle χ_1 is defined as the C–CA–CB–CG dihedral, consistent with the CHARMM force field.²² During the umbrella sampling simulations, the motions of the backbones were restrained by harmonic potentials around the conformation observed in the protein. The umbrella sampling simulations were unbiased with the weighted histogram analysis method (WHAM).²³ The resulting PMF along the dihedral angles was then fitted to a cosine Fourier series of the form

$$V(\phi) = \sum_{n=1}^3 K_n (1 + \cos(n(\phi - \phi_{\text{min}}))) \quad (4)$$

The fitting parameters K_n are given in Table 1. Figure 3 shows the PMF and the result of the fit (the black curve is the PMF, and the red curve is the fitted cosine Fourier series). An appropriate boosting potential can easily be constructed by inverting the sign of the PMF $V(\phi)$ in eq 4 via the CONS DIHE command of CHARMM,¹⁹ thereby canceling the potential barrier between the rotamers of a side chain for any selected residues.

C. MD Simulations. All the FEP/REMD simulations for the binding site were carried out on the IBM Blue Gene/P cluster Intrepid of the Argonne Leadership Computing Facility (ALCF) at Argonne National Laboratory (ANL). The simulations were carried out in a high performance mode using version c36a2 of the CHARMM program,¹⁹ which was

a) FEP/ λ -REMD scheme

b) FEP/H-REMD scheme

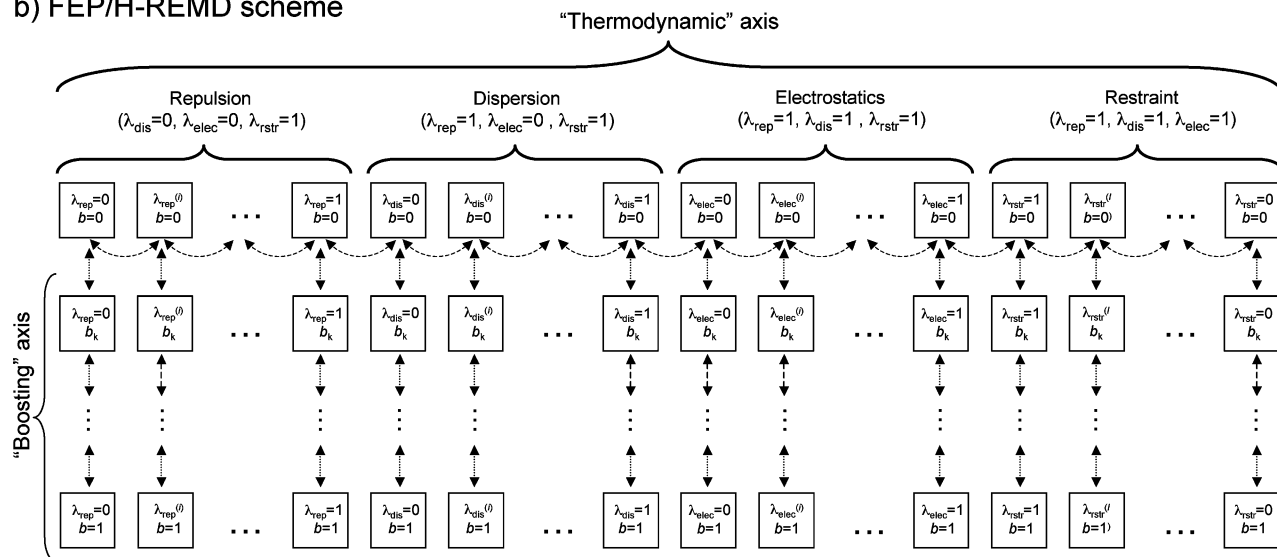


Figure 2. REPDSTR implementation of replica-exchange FEP simulation protocol in the context of the reversible work staging process for ligand binding free energy computations. Each square box represents an atomic simulation with its own I/O. Panel a illustrates the FEP/REMD scheme along the axis of the reversible thermodynamic work with coupling parameter λ (“ λ -swap” moves).¹⁵ The curly dashed-line arrows indicate the possible attempted exchanges, which are allowed only between neighboring replicas along the reversible staging process. At each cycle, the trial exchanges alternate between odd and even numbered replicas (ranked from 1 to N_{tot}), where even exchange means between windows 0 and 1, 2 and 3, 4 and 5, etc. and odd exchange means between windows 1 and 2, 3 and 4, 5 and 6, etc. Panel b illustrates the FEP/H-REMD scheme, where a vertical branch of M boosting-biasing replicas is attached to each of the windows along the reversible work process shown in panel a. The possible attempted moves, indicated by the dashed-line arrows, are again only allowed between neighboring replicas. In the FEP/H-REMD scheme, replica exchange is alternatively attempted along the axis of the reversible thermodynamic work with coupling parameter λ (curly horizontal arrows), and along the biasing axis with boosting parameter b (straight vertical arrows). Each exchange cycle consists of four stages: even and odd *local* exchanges between the biasing replicas within a host FEP window and even and odd *global* exchanges between those FEP windows with $b = 0$.

Table 1. Fitting Parameters of χ_1 PMF with Cosinus-Fourier Series

residue	K_3 (kcal/mol)	ϕ_{\min} (deg)	n	K_2 (kcal/mol)	ϕ_{\min} (deg)	n	K_1 (kcal/mol)	ϕ_{\min} (deg)	n
Val	2.9873	118.86°	3	-0.7662	78.96	2	0.7360	25.18	1
Ile	2.9407	119.07°	3	-0.7178	96.73	2	0.9651	31.86	1
Leu	2.4730	117.59	3	-0.7547	68.32	2	1.4487	-17.81	1
Tyr	2.3712	113.03	3	-0.7047	77.13	2	1.2107	-6.354	1

modified and extended for the present study. The hydration computations were performed on the IBM quads computing cluster KBT at ANL. The binding site free energy simulations were carried out on a reduced model of a solvated T4L/L99A system with the generalized solvent boundary potential (GSBP).²⁴ The initial T4L/L99A system was constructed from the crystallographic structure (PDB 187L) as described previously.¹ The hydration free energy computations of the isolated ligands were carried out under PBC conditions at constant pressure. The systems were propagated with a 2 fs time step using Langevin dynamics at a temperature of 298.15 K. For the binding free energy calculation of *p*-xylene, 100

ps production runs were performed for the binding site with a replica-exchange frequency of 1/100 steps. $M = 7$ boosting replicas were used with H-REMD. The force field parameters and initial structure of *p*-xylene were taken from our previous study for the sake of consistency.¹

In all calculations, the energies were collected during the production run and postprocessed using WHAM.²³ For the binding site simulations, the WHAM postprocessing is only applied to replicas with zero biasing potential. To monitor the convergence of the binding site calculation, 20 independent FEP calculations (20×100 ps) were performed consecutively for each system, each starting from the configuration saved at the

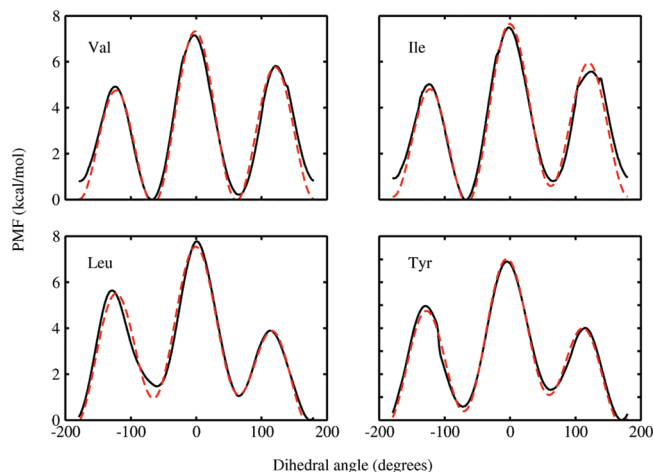


Figure 3. Fitting of χ_1 potential of mean force with linear superposition of three dihedral potential functions. Solid curve, PMF obtained with umbrellas sampling in gas phase; dashed curve, fitting curves.

end of the previous run. The last 10 runs were used to produce the final averaged result and calculate the standard deviations.

Results and Discussion

Both crystallographic studies and computations indicate that the side chain of Val111 of T4L/L99A changes its rotameric states when moderately large ligands bind to the nonpolar cavity.^{1,25} In the case of *p*-xylene, the side chain of Val111 rotates by approximately 140°, to a χ_1 value of -35° , to avoid a steric clash with the ligand (Val111 is depicted in red in Figure 1). It is often convenient to utilize the *holo* configuration as the starting set of coordinates in absolute binding free energy calculations since this is provided either by the X-ray crystallographic structure of the bound complex or by the output of *in silico* docking. However, one must ensure that FEP/MD simulations do reversibly cover the relevant set of thermodynamic states. In principle, an ideal sampling should be able to reflect any conformational changes within the protein between the two end-point *holo* and *apo* states along the thermodynamic decoupling simulations. However, as shown in Figure 4, no spontaneous dihedral transition is observed in simple FEP/REMD simulations. The side chain remains kinetically “trapped” in its *holo* rotameric state, with χ_1 around -60° . This is consistent with the observation reported by Dill and co-workers; the dihedral of the Val111 is unable to cross spontaneously the energy barrier during the FEP simulations.⁴ In previous calculation performed by Deng and Roux with FEP/MD simulations, a 300 ps production run started from the *holo* state resulted in a binding free energy of -9.06 kcal/mol, which is considerably overestimated when compared to the experimental value of -4.7 kcal/mol. A straight FEP/REMD scheme, by itself, improves the value to -6.4 kcal/mol (Table 2). However, the result remains too favorable compared to the experiment.

To address the issue of a kinetically trapped degree of freedom and to enhance the sampling of rotameric states, the extended FEP/H-REMD framework is introduced. As a preliminary test, the boosting potential in FEP/H-REMD was applied exclusively to the χ_1 degree of freedom of Val111, which is the most problematic residue. Eight biasing replicas were used to guarantee a high acceptance ratio (>80%) for exchange

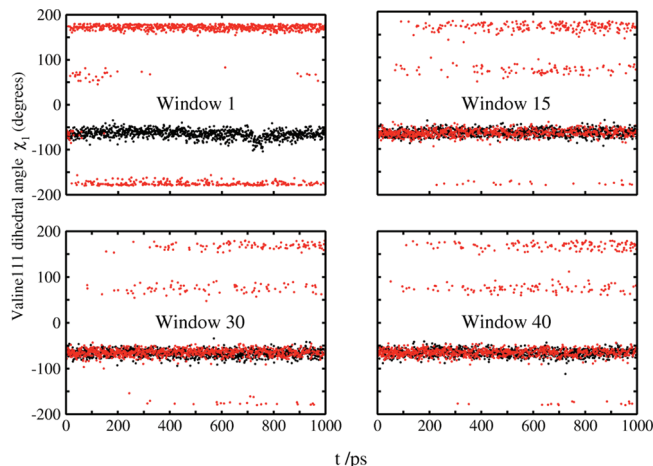


Figure 4. Enhanced sampling of rotameric states of Valine111. Black dots, values obtained with FEP/REMD; red dots, obtained with FEP/H-REMD. The investigated dihedral is χ_1 , C–CA–CB–CG2 dihedral in the CHARMM force field. Window #1 turns off all ligand–receptor interactions and therefore corresponds to the *apo* state. Window #40 turns on all ligand–receptor interactions and corresponds to the *apo* state. With an increasing window index from 1 to 40, the thermodynamic state changes progressively from *apo* to *holo*.

attempts between the replicas with adjacent values of the boosting parameter b . A binding free energy of -5.1 kcal/mol is achieved using this FEP/H-REMD scheme (Table 2), closer to the experimental value, and also in good agreement with the value of -5.06 kcal/mol obtained by Dill and co-workers using a PMF-based confine-and-release strategy.⁴ The largest change occurs for the repulsive free energy contribution ΔG_{rep} , which increases by about 1.5 kcal/mol. The changes for the dispersive and electrostatic free energy contributions are smaller and do not suffer from convergence problems, most likely because they are switched-on after the repulsive core of the ligand has been inserted into the binding cavity. The enhancement of the conformational sampling for the calculation of the repulsive free energy contribution suggests that the dihedral energy barrier is associated with a steric contact with the ligand.

Further insight can be obtained by considering the time evolution of the χ_1 dihedral of Val111, as illustrated in Figure 4. Because all the windows of the FEP/H-REMD simulations are started with Val111 in the rotamer taken from the *holo* state X-ray structure (χ_1 near -60°), the time evolution of the χ_1 dihedral for the *apo* state (i.e., the first window with a completely decoupled ligand) is of particular interest. In the FEP/H-REMD simulations, it is observed that the χ_1 of Val111 rapidly starts to make transitions within ~ 10 ps toward 180° (red curve), corresponding to the dominant rotamer for the *apo* state. In contrast, the side chain remains trapped, with χ_1 around -60° in the straight FEP/REMD simulations (black curve). Moreover, the time evolution of χ_1 for the *holo* state (window #40) fluctuates predominantly around 180° , with some excursions to other values. Along the thermodynamic coupling axis (λ), the population of rotamers changes progressively from the appropriate distribution of the *apo* and *holo* states. For the *apo* state, the average populations for *trans*, *gauche*⁺, and *gauche*⁻ are 0.99, 0.01, and 0.0, respectively. These results are in good

Table 2. Absolute Binding Free Energies of *p*-Xylene to T4L/L99A (all values in kcal/mol)

	binding site			bulk water	exp
	REMD	H-REMD (Val111)	H-REMD (7 residues)	PBC	
ΔG_{rep}	12.00 ± 0.21	13.55 ± 0.17	13.71 ± 0.10	16.02 ± 0.27	
ΔG_{disp}	-25.08 ± 0.07	-25.25 ± 0.08	-25.25 ± 0.08	-15.44 ± 0.02	
ΔG_{elec}	-0.74 ± 0.02	-0.78 ± 0.01	-0.78 ± 0.01	-1.61 ± 0.02	
ΔG_{rstr}	7.02 ± 0.09	7.02 ± 0.09	7.02 ± 0.09		
total	-7.45 ± 0.23	-6.12 ± 0.19	-5.96 ± 0.17	-1.03 ± 0.24	-0.87
$\Delta G_{\text{b}}^{\circ}$	-6.42 ± 0.21	-5.09 ± 0.20	-4.93 ± 0.18		-4.67

accord with the values estimated from the PMF of Dill and co-workers (0.99, 0.01, and 0.0008).⁴ For the *holo* state, the average populations are 0.16, 0.11, and 0.73, again in good accord with the values estimated from the PMF reported by Dill and co-workers (0.23, 0.002, and 0.76). For both the *apo* and *holo* states, the three possible rotamers are ranked correctly, and the probability of the dominant state is reproduced within a few percent. The enhanced sampling provided by FEP/H-REMD is key to producing an accurate binding free energy started from the *holo* configuration.

The ultimate aim of the FEP/H-REMD framework is to enable binding free energy calculation without any prior knowledge of the location of a possible high potential barrier, such as the χ_1 of Val111 in T4L/L99A. In a next round of FEP/H-REMD calculations, we test this concept by applying indiscriminately a biasing boosting potential to seven protein side chains within a distance of 6 Å around the ligand. The residues affected by the boosting potentials are Leu84, Leu91, Leu118, Leu121, Val111, Ile78, and Tyr88. The affected residues around the binding pocket are displayed in color in Figure 1. It should be noted that the selection of residues is done on the basis of the starting (*holo*) configuration and remains unchanged during the entire FEP/H-REMD calculation. In this illustrative test, the stages corresponding to the dispersive and charging contributions were skipped in order to focus mainly on the dominant repulsive contribution. The results given in Table 2 show that a binding free energy of -4.9 kcal/mol is obtained, essentially identical with the calculation where only χ_1 of Val111 was boosted. Figure 5 shows the sampling of rotameric states of four selected residues about the binding pocket. For this FEP/H-REMD calculation, simple boosting potentials extracted from the PMF of small peptides in the gas phase were used. More sophisticated constructs could certainly be designed to further improve the quality of the boost potential. For example, one could also switch off the nonbonding between different side chains to resolve the problem of a kinetic bottleneck caused by steric clashes. Nevertheless, a simple PMF bias seems to be sufficient to enhance the sampling in the present application. For Val111, the distribution of the χ_1 value remains the same as for the former calculations. For Leu and Tyr, no interconversion among rotameric states is observed, which is consistent with the *apo* and *holo* X-ray structures. Interestingly, some infrequent transitions between two rotameric states of Ile78 are also observed. While the present application concerns the binding of *p*-xylene to a relatively small nonpolar cavity, the FEP/H-REMD scheme is expected to scale efficiently with an increased number of freedoms.

It should be emphasized that the enhanced sampling is achieved without any prior knowledge of all the relevant degrees

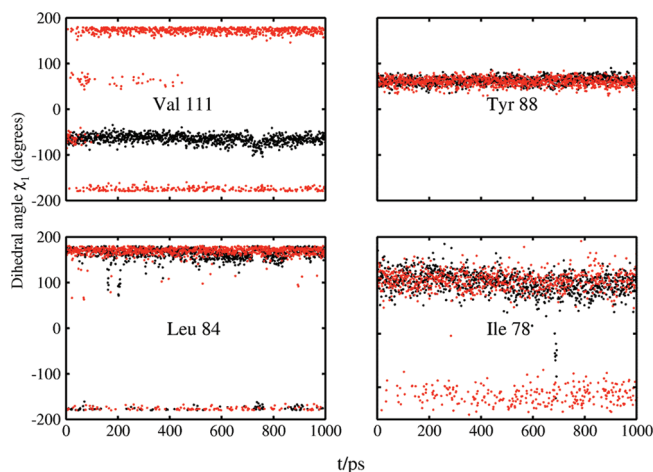


Figure 5. Rotameric states of four selected residues closest to the ligand. The four panels show the χ_1 of residues Val111, Tyr88, Leu84, and Ile78 in the *apo* state. It should be noted that Ile78 obtains a considerable sampling enhancement with FEP/H-REMD.

of freedom that are kinetically trapped. Nevertheless, it is necessary to apply the boosting potential to a finite subset of side chains and treat those via H-REMD. An important advantage is that the FEP/H-REMD scheme is not expected to be strongly size-limited. In principle, the proposed framework is applicable to a region of interest around the ligand, without a considerable loss in efficiency. As the number of side chains grows, longer simulations could be needed to get the same level of sampling. However, the moderate size of typical drug-like molecules ensures that only a finite and relatively small number of side chains should be treated with H-REMD. Effective boost potentials can be precalculated and stored in a library for any type of residues to achieve a universal sampling enhancement of side chain rotamers in the neighborhood of any binding pockets. Extensions to the present framework to include backbone degrees of freedom are in progress.

Conclusion

In summary, a dual FEP simulation scheme suitable for a large supercomputing platform was proposed to enhance the sampling of protein side chains in binding free energy calculations. Extending from our previous work,¹⁵ each system with a given thermodynamic coupling factor λ in the extended ensemble in FEP/REMD is further coupled with a set of replicas evolving on a biased energy surface with boosting potentials accelerating the interconversion among different rotameric states of a set of side chains in the neighborhood of the binding site via a Hamiltonian REMD scheme. An important feature of the FEP/H-REMD scheme is that it can be used to enhance the sampling

of a fairly large number of putative slowly varying degrees of freedom without a considerable loss in efficiency. Sampling of any residue lining the binding pocket can benefit by the boosting H-REMD from a set of precalculated biasing potentials stored in a library. Application of FEP/H-REMD shows that the sampling of rotamers of the side chains surrounding the nonpolar cavity of T4L/L99A is significantly enhanced and that the binding free energy for a large ligand such as *p*-xylene can be calculated accurately by starting from the *holo* protein configuration. Further developments of the present method to include the treatment of backbone reorganization are currently in progress.

Acknowledgment. We are grateful to Dr. Andrew Binkowski for his support. We would like to acknowledge Dr. Milan Hodoscek for collaboration with the programming work for the CHARMM REPDSTR module, and Dr. Albert Lau for valuable discussions about free energy calculations and the replica-exchange scheme. This research is funded by grant MCB-0920261 from the National Science Foundation. Access to the computational resources of ALCF at ANL, supported by the Office of Science of the U.S. Department of Energy (DOE) under contract DE-AC02-06CH11357, was made possible by an INCITE grant from the DOE. The submitted manuscript has been created by UChicago Argonne, LLC, Operator of ANL. ANL, a U.S. DOE Office of Science laboratory, is operated under Contract No. DE-AC02-06CH11357.

References

- (1) Deng, Y.; Roux, B. *J. Chem. Theory Comput.* **2006**, *2*, 1255–1273.
- (2) Wang, J.; Deng, Y.; Roux, B. *Biophys. J.* **2006**, *91*, 2798–2814.
- (3) Mobley, D. L.; Graves, A. P.; Chodera, J. D.; McReynolds, A. C.; Shoichet, B. K.; Dill, K. A. *J. Mol. Biol.* **2007**, *371*, 1118–1134.
- (4) Mobley, D. L.; Chodera, J. D.; Dill, K. A. *J. Chem. Theory Comput.* **2007**, *3*, 1231–1235.
- (5) Woo, H. J.; Roux, B. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6825–6830.
- (6) Gan, W.; Roux, B. *Proteins* **2009**, *74*, 996–1007.
- (7) Hritz, J.; Oostenbrink, C. *J. Phys. Chem. B* **2009**, *113*, 12711–20.
- (8) Jorgensen, W. *Acc. Chem. Res.* **2009**, *42*, 724.
- (9) Hamelberg, D.; Mongan, J.; McCammon, J. A. *J. Chem. Phys.* **2004**, *120*, 11919–11929.
- (10) Kannan, S.; Zacharis, M. *Proteins: Struct., Funct., Bioinf.* **2007**, *66*, 697–706.
- (11) Chao, X.; Wang, J.; Liu, H. *J. Chem. Theory Comput.* **2008**, *4*, 1348–1359.
- (12) Straatsma, T. P.; McCammon, J., A. *J. Chem. Phys.* **1994**, *101*, 5032–5039.
- (13) Fajer, M.; Hamelberg, D.; McCammon, J. A. *J. Chem. Theory Comput.* **2008**, *4*, 1565–1569.
- (14) Hritz, J.; Oostenbrink, C. *J. Chem. Phys.* **2008**, *128*, 144121.
- (15) Jiang, W.; Hodoscek, M.; Roux, B. *J. Chem. Theory Comput.* **2009**, *5*, 2583–2588.
- (16) Woods, C. J.; Essex, J. W.; King, M. A. *J. Phys. Chem. B* **2003**, *107*, 13711–13718.
- (17) Woodcock, H. L., III; Hodoscek, M.; Sherwood, P.; Lee, Y.; Schaefer, H.; Brooks, B. *Theor. Chem. Acc.* **2003**, *109*, 140–148.
- (18) Woodcock, H. L., III; Hodoscek, M.; Gilbert, A. T. B.; Gill, P. M. W.; Schaefer, H. F., III; R., B. B. *J. Comput. Chem.* **2007**, *28*, 1485–1502.
- (19) Brooks, B. R.; Brooks, C. L., III; Mackerell, A. D., Jr.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; Caffisch, A.; Caves, L.; Cui, Q.; Dinner, A. R.; Feig, M.; Fischer, S.; Gao, J.; Hodoscek, M.; Im, W.; Kuczera, K.; Lazaridis, T.; Ma, J.; Ovchinnikov, V.; Paci, E.; Pastor, R. W.; Post, C. B.; Pu, J. Z.; Schaefer, M.; Tidor, B.; Venable, R. M.; Woodcock, H. L.; Wu, X.; Yang, W.; York, D. M.; Karplus, M. *J. Comput. Chem.* **2009**, *30*, 1545–1614.
- (20) Deng, Y.; Roux, B. *J. Phys. Chem.* **2004**, *108*, 16567–16576.
- (21) Weeks, J. D.; Chandler, D.; Anderson, H. C. *J. Chem. Phys.* **1971**, *54*, 5237–5247.
- (22) MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Wantanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M. *J. Phys. Chem. B* **1998**, *102*, 3586–3616.
- (23) Kumar, S.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A.; Rosenberg, J. M. *J. Comput. Chem.* **1992**, *13*, 1011–1021.
- (24) Im, W.; Berneche, S.; Roux, B. *J. Chem. Phys.* **2000**, *114*, 2924–2937.
- (25) Morton, A.; Matthews, B. W. *Biochemistry* **1995**, *34*, 8576–8588.

CT1001768

JCTC

Journal of Chemical Theory and Computation

Proton Transfer Studied Using a Combined Ab Initio Reactive Potential Energy Surface with Quantum Path Integral Methodology

Kim F. Wong,^{†,∇} Jason L. Sonnenberg,^{‡,○} Francesco Paesani,^{†,◆} Takeshi Yamamoto,[§]
 Jiří Vaníček,^{||,¶} Wei Zhang,[⊥] H. Bernhard Schlegel,^{*,‡} David A. Case,[⊥]
 Thomas E. Cheatham III,[#] William H. Miller,^{*,||} and Gregory A. Voth^{*,†,⊗}

Center for Biophysical Modeling and Simulations and Department of Chemistry, University of Utah, Salt Lake City, Utah 84112, Department of Chemistry, Wayne State University, Detroit, Michigan 48202, Department of Chemistry, Graduate School of Science, Kyoto University, Kyoto 606-8502, Japan, Department of Chemistry and Kenneth S. Pitzer Center for Theoretical Chemistry, and Chemical Science Division, Lawrence Berkeley National Laboratory, University of California, Berkeley, California 94720, BioMaPS Institute and Department of Chemistry & Chemical Biology, Rutgers University, Piscataway, New Jersey 08854, and Departments of Medicinal Chemistry, Pharmaceuticals and Pharmaceutical Chemistry, and Bioengineering, 2000 East, 30 South, Skaggs Hall 201, University of Utah, Salt Lake City, Utah 84112

Received November 2, 2009

Abstract: The rates of intramolecular proton transfer are calculated on a full-dimensional reactive electronic potential energy surface that incorporates high-level ab initio calculations along the reaction path and by using classical transition state theory, path-integral quantum transition state theory, and the quantum instanton approach. The specific example problem studied is malonaldehyde. Estimates of the kinetic isotope effect using the latter two methods are found to be in reasonable agreement with each other. Improvements and extensions of this practical, yet chemically accurate framework for the calculations of quantized, reactive dynamics are also discussed.

1. Introduction

The breaking and formation of chemical bonds is fundamental to chemistry. While molecular dynamics (MD)

simulations of large-scale assemblies for hundreds of nanoseconds and even for a few microseconds are permissible using current terascale and emergent petascale high-performance computing infrastructures,^{1–3} the majority of these scientific applications involving conformational sampling or molecular association processes cannot model chemical reactions. Much of current force field research

* Corresponding authors. Phone: 313-577-2562 (H.B.S.), 510-642-0653 (W.H.M.), 773-702-9092 (G.A.V.). Fax: 313-577-8822 (H.B.S.), 510-642-6262 (W.H.M.), 773-795-9106 (G.A.V.). E-mail: hbs@chem.wayne.edu (H.B.S.), millerwh@berkeley.edu (W.H.M.), gavoth@uchicago.edu (G.A.V.).

[†] Center for Biophysical Modeling and Simulations and Department of Chemistry, University of Utah.

[‡] Wayne State University.

[§] Kyoto University.

^{||} University of California, Berkeley.

[⊥] Rutgers University.

[#] Departments of Medicinal Chemistry, Pharmaceuticals and Pharmaceutical Chemistry, and Bioengineering, University of Utah.

[∇] Current Address: Center for Simulation and Modeling, 205 Bellefield Hall, University of Pittsburgh, Pittsburgh, Pennsylvania 15213.

[○] Current Address: Department of Chemistry, Stevenson University, 1525 Greenspring Valley Road, Stevenson, Maryland, 21153.

[◆] Current Address: Department of Chemistry and Biochemistry, University of California, San Diego, 9500 Gilman Drive, La Jolla, California 92093.

[¶] Current Address: Ecole polytechnique fédérale de Lausanne, Institut des sciences et ingénierie chimiques, EPFL SB ISIC LCPT, BCH 3110 (Bât. BCH), CH-1015 Lausanne, Switzerland.

[⊗] Current Address: Department of Chemistry, The University of Chicago, 5735 S. Ellis Avenue, Chicago, Illinois 60637.

focuses on the development of accurate and reliable interactions for describing protein structures, solvation energies, and hybrid bioinorganic material properties.^{4–10} Designing a chemically reactive potential energy surface (PES), nevertheless, is not a new idea.¹¹ During the early half of the twentieth century, chemical physics pioneers, such as Eyring¹² and Evans and Polanyi,^{13–15} proposed procedures for forming surfaces to describe diverse systems involving ionic, S_N2 and Diels–Alder reactions. The common element among these methodologies is the description of the system within a matrix representation for the Hamiltonian:

$$\hat{H} = \begin{bmatrix} H_{11} & \cdots & H_{1N} \\ \vdots & \ddots & \vdots \\ H_{N1} & \cdots & H_{NN} \end{bmatrix} \quad (1)$$

$$H\Psi = \varepsilon\Psi \quad (2)$$

Here, the system evolves on the lowest energy eigenstate, ε_0 , of the Hamiltonian as a superposition state of multiple nuclear-electronic configurations.¹⁶ Variations of this technique differ primarily in the treatments of the H_{ii} and H_{ij} terms. Notably, Warshel and Weiss utilized a combination of molecular dynamics (MD) force fields and empirical fitting, an empirical valence bond (EVB) approach, to approximate the matrix elements for describing reactions in solutions and within enzyme environments.^{17,18}

Although the empirical approach is a computationally practical strategy for modeling chemistry, a general framework for constructing reactive potential energy surfaces based on *first principles*, ab initio, information is desirable. Empirical procedures typically involve a high startup cost in the form of parameter fitting and calibrations that in some instances may be system-specific. On the other hand, even with current advances in computation infrastructure (both from an algorithms and from a hardware perspective), direct ab initio MD approaches are limited from small to moderate sized systems for a few hundred picoseconds of conformational sampling.^{19–24} These time scales cannot capture the physics of many biologically relevant reactions, typically occurring in the microseconds or on longer time scales. Furthermore, the computation time required for adequately sampling phase space to obtain converged thermodynamic observables precludes the use of direct ab initio MD approaches for complex systems consisting of 10 000 atoms or more. A general hybrid simulation framework combining the computational advantages of MD simulation with the accuracy of electronic structure methods is appealing.

In the present paper, we describe an implementation of such a general framework, the distributed Gaussian (DG) approach,^{25–27} for developing accurate ab initio-based potential energy surfaces capable of modeling chemical reactions. The purpose of this paper is to describe the integration, within the Amber²⁸ biosimulations suite, of the DG-EVB surface generation methodology with molecular dynamics, path integrals,^{29–31} and quantum instanton^{32–35} approaches for the computation of thermal rate constants. This methodology development is tested on a well-established system, the prototype intramolecular proton transfer reaction in malonaldehyde coupled to a thermal bath. Despite the

presumed simplicity of malonaldehyde, it is only recently that full-dimensional quantum calculations of tunneling splitting based on accurate ab initio surfaces were possible.^{36–40} Previous works utilized reduced-dimensional approximations or less accurate potentials that may not fully capture the physics of the reaction.^{41–47} Malonaldehyde thus provides a well-studied yet challenging test for assessing the diverse components of the DG-EVB methodology. Integrating both molecular mechanical and ab initio elements, DG-EVB is similar in strategy to Truhlar's recently developed multi-configuration molecular mechanics (MCMM) method.^{48–52}

The remainder of this paper is organized as follows: To show the relationship between the various methods, the theoretical backgrounds for the DG method, the path integral approach and the QI method are briefly reviewed. The simulation details are described in section 3 followed by the Results and Discussion in section 4. Last, we conclude with an assessment of the method and suggestions for future improvements.

2. Theoretical Background

This section briefly describes the theoretical basis to our algorithmic development for constructing a reactive PES from ab initio information and the incorporation of nuclear quantum effects in molecular dynamics. Dynamical methods for computing thermal rate constants and kinetic isotope effects are described. Our test system of a single malonaldehyde molecule coupled to a thermal reservoir is modeled within the canonical ensemble. Within the canonical ensemble, thermal rate constants are well-defined quantities.⁵³ Under favorable circumstances, tunneling splittings can also be extracted from the imaginary time correlation function from PIMD simulations, but in the present case reliable error bounds on the tunneling splitting could not be obtained using the maximum entropy approach.

2.1. Designing ab Initio Based Chemically Reactive PESs. A number of methods for building an ab initio derived PES using eq 1 has been proposed recently. Among these are Chang–Miller,^{54,55} Minichino–Voth,⁵⁶ multiconfiguration molecular mechanics (MCMM),^{48–52} and our distributed Gaussian EVB (DG-EVB).^{25–27} Since the distributed Gaussian approach is related to the Chang–Miller prescription, we concisely describe Chang–Miller below to motivate the subsequent method development. The Chang–Miller approach attempts to construct an accurate reactive potential energy surface by fitting a superposition of reactant and product configurations using a generalized Gaussian form for the coupling term

$$H_{12}^2(\mathbf{q}) = A \exp[\mathbf{B}^T \cdot \Delta\mathbf{q} - 1/2 \Delta\mathbf{q}^T \cdot \tilde{\mathbf{C}} \cdot \Delta\mathbf{q}], \Delta\mathbf{q} = \mathbf{q} - \mathbf{q}_{TS} \quad (3)$$

The diagonal elements describing the reactant state (RS) and product state (PS) are approximated using conventional, nonreactive force fields (FF). This choice was motivated by the expectation that the development of classical force fields will continually improve toward a state where they are accurate *enough*. Formulated to reproduce the ab initio

energy ε_Ψ , gradient, and Hessian at the transition state geometry, \mathbf{q}_{TS} , the parameters A (a scalar), \mathbf{B} (a vector), and $\tilde{\mathbf{C}}$ (a matrix) take the analytical forms

$$A = [H_{11}(\mathbf{q}_{\text{TS}}) - \varepsilon_\Psi(\mathbf{q}_{\text{TS}})][H_{22}(\mathbf{q}_{\text{TS}}) - \varepsilon_\Psi(\mathbf{q}_{\text{TS}})] \quad (4)$$

$$\mathbf{B} = \frac{\mathbf{G}_1}{[H_{11}(\mathbf{q}_{\text{TS}}) - \varepsilon_\Psi(\mathbf{q}_{\text{TS}})]} + \frac{\mathbf{G}_2}{[H_{22}(\mathbf{q}_{\text{TS}}) - \varepsilon_\Psi(\mathbf{q}_{\text{TS}})]} \quad (5)$$

$$\mathbf{G}_N = \left. \frac{\partial H_{NN}(\mathbf{q})}{\partial \mathbf{q}} \right|_{\mathbf{q}=\mathbf{q}_{\text{TS}}} - \left. \frac{\partial \varepsilon_\Psi(\mathbf{q})}{\partial \mathbf{q}} \right|_{\mathbf{q}=\mathbf{q}_{\text{TS}}}$$

$$\tilde{\mathbf{C}} = \frac{\mathbf{G}_1 \mathbf{G}_1^T}{[H_{11}(\mathbf{q}_{\text{TS}}) - \varepsilon_\Psi(\mathbf{q}_{\text{TS}})]} + \frac{\mathbf{G}_2 \mathbf{G}_2^T}{[H_{22}(\mathbf{q}_{\text{TS}}) - \varepsilon_\Psi(\mathbf{q}_{\text{TS}})]} - \frac{\tilde{\mathbf{K}}_1}{[H_{11}(\mathbf{q}_{\text{TS}}) - \varepsilon_\Psi(\mathbf{q}_{\text{TS}})]} - \frac{\tilde{\mathbf{K}}_2}{[H_{22}(\mathbf{q}_{\text{TS}}) - \varepsilon_\Psi(\mathbf{q}_{\text{TS}})]} \quad (6)$$

$$\tilde{\mathbf{K}}_N = \left. \frac{\partial^2 H_{NN}(\mathbf{q})}{\partial \mathbf{q}^2} \right|_{\mathbf{q}=\mathbf{q}_{\text{TS}}} - \left. \frac{\partial^2 \varepsilon_\Psi(\mathbf{q})}{\partial \mathbf{q}^2} \right|_{\mathbf{q}=\mathbf{q}_{\text{TS}}}$$

When the system configuration deviates far from the transition state structure (i.e., for large $\Delta \mathbf{q}$) and the matrix $\tilde{\mathbf{C}}$ contains one or more negative frequencies, H_{12}^2 diverges. Although refinements are available for controlling the asymptotic behavior of the Chang–Miller approach,⁵⁵ simply recasting eq 3 in terms of a quadratic polynomial times a spherical Gaussian

$$H_{12}^2(\mathbf{q}) = A[1 + \mathbf{B}^T \cdot \Delta \mathbf{q} + \frac{1}{2} \Delta \mathbf{q}^T \cdot (\tilde{\mathbf{C}}' + \alpha \tilde{\mathbf{I}}) \cdot \Delta \mathbf{q}] \times \exp[-\frac{1}{2} \alpha |\Delta \mathbf{q}|^2] \quad (7)$$

keeps the coupling element bounded at the asymptotes.²⁵ The scalar A and vector \mathbf{B} parameters are identical to those in the Chang–Miller approach, while the matrix assumes a modified form

$$\tilde{\mathbf{C}}' = \frac{\mathbf{G}_1 \mathbf{G}_1^T + \mathbf{G}_2 \mathbf{G}_2^T}{A} + \frac{\tilde{\mathbf{K}}_1}{[H_{11}(\mathbf{q}_{\text{TS}}) - \varepsilon_\Psi(\mathbf{q}_{\text{TS}})]} + \frac{\tilde{\mathbf{K}}_2}{[H_{22}(\mathbf{q}_{\text{TS}}) - \varepsilon_\Psi(\mathbf{q}_{\text{TS}})]} \quad (8)$$

Note that $\tilde{\mathbf{I}}$ is the identity matrix and α is a parameter related to the Gaussian width. The distributed Gaussian approach generalizes the above polynomial times a Gaussian prescription [eq 7] to utilize ab initio information not only at the transition state geometry but also at other points on the potential energy surface. Here, $H_{12}^2(\mathbf{q})$ is approximated as an expansion in a set of *distributed Gaussians* centered on a set of molecular configurations \mathbf{q}_K :

$$H_{12}^2(\mathbf{q}) = \sum_K^{N_{\text{cfg}}} \sum_{i,j \geq 0}^{N_{\text{dim}}} B_{ijk} g(\mathbf{q}, \mathbf{q}_K, i, j, \alpha_K) \quad (9)$$

$$g(\mathbf{q}, \mathbf{q}_K, 0, 0, \alpha_K) = (1 + \frac{1}{2} \alpha_K |\Delta \mathbf{q}_K|^2) \exp[-\frac{1}{2} \alpha_K |\Delta \mathbf{q}_K|^2] \quad (10)$$

$$g(\mathbf{q}, \mathbf{q}_K, i, 0, \alpha_K) = (\Delta \mathbf{q}_K)_i \exp[-\frac{1}{2} \alpha_K |\Delta \mathbf{q}_K|^2] \quad (11)$$

$$g(\mathbf{q}, \mathbf{q}_K, i, j, \alpha_K) = (1 - \frac{1}{2} \delta_{ij}) (\Delta \mathbf{q}_K)_i (\Delta \mathbf{q}_K)_j \times \exp[-\frac{1}{2} \alpha_K |\Delta \mathbf{q}_K|^2] \quad (12)$$

$$\Delta \mathbf{q}_K = \mathbf{q} - \mathbf{q}_K \quad (13)$$

where N_{cfg} is the number of ab initio data points used for the fitting; N_{dim} is the number of system coordinates; $g(\mathbf{q}, \mathbf{q}_K, i, j, \alpha_K)$ are the s-, p- and d-type Gaussians; and B_{ijk} are the expansion coefficients. The term involving the identity matrix in eq 7 was accumulated into the s-type Gaussian [see eq 10] to precondition the system of linear equations for efficient convergence when utilizing iterative methods. The nonstandard form of the d-type Gaussian is for similar reasons. If the number of Gaussian centers, K , is equal to the number of data points where $H_{12}^2(\mathbf{q})$ is evaluated, eq 9 describes a system of linear equations

$$\mathbf{F} = \tilde{\mathbf{D}} \mathbf{B} \quad (14)$$

that can be solved using singular value decomposition or by an iterative procedure, such as GMRES (generalized minimal residual method).^{57–60} When derivatives for the coupling terms are available, this information can also be utilized for the fitting, i.e.,

$$\left. \frac{\partial H_{12}^2(\mathbf{q})}{\partial \mathbf{q}} \right|_{\mathbf{q}=\mathbf{q}_L} = \sum_K^{N_{\text{cfg}}} \sum_{i,j \geq 0}^{N_{\text{dim}}} B_{ijk} \left. \frac{\partial g(\mathbf{q}, \mathbf{q}_K, i, j, \alpha_K)}{\partial \mathbf{q}} \right|_{\mathbf{q}=\mathbf{q}_L} \quad (15)$$

$$\left. \frac{\partial^2 H_{12}^2(\mathbf{q})}{\partial \mathbf{q}^2} \right|_{\mathbf{q}=\mathbf{q}_L} = \sum_K^{N_{\text{cfg}}} \sum_{i,j \geq 0}^{N_{\text{dim}}} B_{ijk} \left. \frac{\partial^2 g(\mathbf{q}, \mathbf{q}_K, i, j, \alpha_K)}{\partial \mathbf{q}^2} \right|_{\mathbf{q}=\mathbf{q}_L} \quad (16)$$

The \mathbf{F} column vector stores the terms $H_{12}^2(\mathbf{q}_L)$, $\partial H_{12}^2(\mathbf{q})/\partial \mathbf{q}|_{\mathbf{q}=\mathbf{q}_L}$, and $\partial^2 H_{12}^2(\mathbf{q})/\partial \mathbf{q}^2|_{\mathbf{q}=\mathbf{q}_L}$ evaluated at the N_{cfg} ab initio configurations. The $\tilde{\mathbf{D}}$ matrix contains the values of the Gaussian bases, $g(\mathbf{q}_L, \mathbf{q}_K, i, j, \alpha_K)$, $\partial g(\mathbf{q}, \mathbf{q}_K, i, j, \alpha_K)/\partial \mathbf{q}|_{\mathbf{q}=\mathbf{q}_L}$, and $\partial^2 g(\mathbf{q}, \mathbf{q}_K, i, j, \alpha_K)/\partial \mathbf{q}^2|_{\mathbf{q}=\mathbf{q}_L}$, evaluated at these same configurations \mathbf{q}_L . The \mathbf{B} column vector contains the set of unknown expansion coefficients being solved for within this linear system of equations. Once this solution vector is obtained, we have a general (\mathbf{q} -dependent) analytical approximation to the coupling [eq 9] and corresponding derivatives [eqs 15 and 16], computed as a linear combination of the distributed Gaussian basis set.

For a symmetric 2×2 Hamiltonian matrix, the analytical expression for the coupling and derivatives are given by

$$H_{12}^2(\mathbf{q}) = [H_{11}(\mathbf{q}) - \varepsilon_0(\mathbf{q})][H_{22}(\mathbf{q}) - \varepsilon_0(\mathbf{q})] \quad (17)$$

$$\frac{\partial H_{12}^2(\mathbf{q})}{\partial \mathbf{q}} = \left[\frac{\partial H_{11}(\mathbf{q})}{\partial \mathbf{q}} - \frac{\partial \varepsilon_0(\mathbf{q})}{\partial \mathbf{q}} \right] \left[\frac{\partial H_{22}(\mathbf{q})}{\partial \mathbf{q}} - \frac{\partial \varepsilon_0(\mathbf{q})}{\partial \mathbf{q}} \right] \quad (18)$$

$$\frac{\partial^2 H_{12}^2(\mathbf{q})}{\partial \mathbf{q}^2} = \left[\frac{\partial^2 H_{11}(\mathbf{q})}{\partial \mathbf{q}^2} - \frac{\partial^2 \varepsilon_0(\mathbf{q})}{\partial \mathbf{q}^2} \right] \left[\frac{\partial^2 H_{22}(\mathbf{q})}{\partial \mathbf{q}^2} - \frac{\partial^2 \varepsilon_0(\mathbf{q})}{\partial \mathbf{q}^2} \right] \quad (19)$$

where ε_0 is the lowest eigenvalue of the matrix and H_{11} and H_{22} are the reactant and product valence bond states. H_{11} and H_{22} can be described by a force field potential (as in the spirit of Chang–Miller⁵⁴) or as a Taylor series expansion about the respective ab initio minimum.²⁵ The key idea is to determine $H_{12}^2(\mathbf{q})$ such that the resulting ε_0 surface approximates the ab initio surface, i.e., $\varepsilon_0 = \varepsilon_\Psi$. For the above two-state system, the coupling and corresponding derivatives can be evaluated directly from eqs 17–19 using the ab initio energies (ε_Ψ), gradients ($\partial \varepsilon_\Psi / \partial \mathbf{q}$), and Hessian ($\partial^2 \varepsilon_\Psi / \partial \mathbf{q}^2$) data in conjunction with the corresponding H_{ii} values and derivatives. This procedure provides coupling values and derivatives for the N_{cfg} discrete ab initio geometries. The DG-EVB method provides a prescription for evaluating the coupling at all coordinates as a linear combination of Gaussians expanded about these ab initio configurations, which may be chosen to be distributed along the IRC (intrinsic reaction coordinate), for example. While an analytical expression of the H_{ij} term for a general multistate $N \times N$ system Hamiltonian does not exist, one can make, nevertheless, the pairwise approximation and estimate the couplings using the two-state expression above for unique ij pairs. In this paper, we focus only on the two-state problem.

2.2. Incorporating Nuclear Quantum Effects. When the reactive process involves particles whose thermal de Broglie wavelength is comparable to the characteristic scale of the potential, nuclear quantum effects such as tunneling and zero-point motion are important for describing the chemical rate.^{61–63} Feynman’s path integral (PI)^{29–31} formalism of quantum mechanics provides a general framework for describing equilibrium (i.e., thermodynamic and structural) properties of a quantum many-body system. Briefly, the quantum partition function for the canonical (NVT) ensemble can be written as

$$Q = \left(\frac{P}{2\pi\hbar^2\beta} \right)^{3NP/2} \prod_{i=1}^N m_i^{3P/2} \int d\mathbf{q}^{(1)} \dots d\mathbf{q}^{(P)} \exp[-\beta\Phi] \quad (20)$$

where

$$\Phi(\mathbf{q}^{(1)}, \dots, \mathbf{q}^{(P)}) = \frac{P}{2\hbar^2\beta^2} \sum_{i=1}^N m_i \sum_{s=1}^P (\mathbf{q}_i^{(s)} - \mathbf{q}_i^{(s+1)})^2 + \frac{1}{P} \sum_{s=1}^P V(\mathbf{q}^{(s)}) \quad (21)$$

In this form, Q is isomorphic to the classical configurational partition function of N ring polymers, each comprised of P particles.⁶⁴ Each particle within the polymer interacts harmonically with neighbors along the cyclic path (first term) and with other particles within the same s -indexed PI slice

via $V(\mathbf{q}^{(s)})/P$ (second term). By introducing a set of Gaussian integrals into eq 20, one sees that the quantum partition function

$$Q = \Lambda \left(\frac{P}{2\pi\hbar^2\beta} \right)^{3NP/2} \prod_{i=1}^N m_i^{3P/2} \int d\mathbf{p}^{(1)} \dots d\mathbf{p}^{(P)} \int d\mathbf{q}^{(1)} \dots d\mathbf{q}^{(P)} \times \exp \left[-\beta \left(\sum_{i=1}^N \frac{p_i^{(s)}}{2\mu_i^{(s)}} + \Phi \right) \right] \quad (22)$$

can be evaluated from molecular dynamics based on the Newtonian equations of motion derived from a fictitious classical Hamiltonian of the form

$$H(\mathbf{p}^{(1)}, \dots, \mathbf{p}^{(P)}, \mathbf{q}^{(1)}, \dots, \mathbf{q}^{(P)}) = \sum_{i=1}^N \sum_{s=1}^P \left(\frac{p_i^{(s)}}{2\mu_i^{(s)}} + \Phi(\mathbf{q}^{(1)}, \dots, \mathbf{q}^{(P)}) \right) \quad (23)$$

This is possible because the inserted fictitious momenta are uncoupled and thus can be integrated analytically such that the prefactor Λ can be defined to give back the correct quantum partition function in eq 20. The purpose of the fictitious Hamiltonian is simply to provide a computationally practical scheme for evaluating the total phase space integral. Finally, to recover Q in the NVT ensemble, we also need to couple the system to a thermostat so as to ensure that the sampled distribution is indeed canonical.⁶⁵

While the path integral formulation is general for any internuclear potential, the results will depend on the accuracy of the electronic surface. In our study, the DG-EVB potential enters into PIMD through the second term of eq 21. The PI results presented here are computed from the normal mode representation of PIMD,⁶⁶ where the thermostat is a set of Nosé–Hoover chains.⁶⁷

2.3. Computing Observables from Molecular Dynamics Trajectories. The computable quantities connecting theory and simulation to experimental measurements are traditionally referred to as *observables*. For a reactive process, the *naturally* measurable quantity is the chemical rate. Transition state theory (TST)^{53,68–70} provides a simple framework for calculating the rate from molecular dynamics trajectories. In TST, the rate along a *reaction path* $\xi(\mathbf{q})$ [synonymous with the *reaction coordinate* (RC)] is given by

$$k_{\text{TST}} = \frac{1}{2} \langle |d\xi/dt| \rangle_{\xi^{\ddagger}} \langle \rho(\xi^{\ddagger}) \rangle \quad (24)$$

where $d\xi/dt$ is the *dynamical frequency factor* and ρ is the normalized density of sampled RC values

$$\langle \rho(\xi) \rangle = \frac{\int d\mathbf{q} \exp[-\beta H(\mathbf{q})] \delta[\xi(\mathbf{q}) - \xi]}{\int d\mathbf{q} \exp[-\beta H(\mathbf{q})] h[\xi^{\ddagger} - \xi(\mathbf{q})]} \quad (25)$$

We note that the *intrinsic reaction coordinate* (minimum-energy reaction path) is one well-defined prescription for $\xi(\mathbf{q})$, although any RC (subject to the limitations of TST) may be employed. In the above equation, $\beta = 1/k_{\text{B}}T$, h is the Heaviside step function, ξ^{\ddagger} is the location of the *dividing surface* partitioning the reactant and product regions, and ξ

is the value of the RC that was sampled during MD described by the system Hamiltonian $H(\mathbf{q})$. The dynamical frequency factor can be estimated by the velocity of a free particle along the RC direction

$$\langle |d\xi/dt| \rangle_{\xi^{\ddagger}} = \left(\frac{2}{\pi\beta} \right)^{1/2} \left\langle \left[\sum_{i=1}^{3N} \frac{1}{m_i} \left(\frac{\partial \xi(\mathbf{q})}{\partial q_i} \right)^2 \right]^{1/2} \right\rangle_{\xi^{\ddagger}} \quad (26)$$

where $\langle \dots \rangle_{\xi^{\ddagger}}$ denotes the conditional average computed at the dividing surface ξ^{\ddagger} . Both the average distribution of RC values in eq 25 and the gradient of the RC in eq 26 are readily computable from MD using umbrella sampling techniques, for example.

For pedagogical purposes, it is convenient to recast the average distribution function

$$\langle \tilde{\rho}(\xi) \rangle = \frac{\int d\mathbf{q} \exp[-\beta H(\mathbf{q})] \delta[\tilde{\xi}(\mathbf{q}) - \xi]}{\int d\mathbf{q} \exp[-\beta H(\mathbf{q})]} \quad (27)$$

in the perspective of the *potential of mean force* (PMF)⁷¹

$$w(\xi) = w(\xi^0) - \beta^{-1} \ln \left[\frac{\langle \tilde{\rho}(\xi) \rangle}{\langle \tilde{\rho}(\xi^0) \rangle} \right] \quad (28)$$

where ξ^0 and $w(\xi^0)$ are constants that are typically chosen to reflect initial conditions of the chemical system. For example, note that $\langle \rho(\xi) \rangle$ and $\langle \tilde{\rho}(\xi) \rangle$ in the above equations differ only by a normalization factor. This factor, within transition state theory, is chosen to normalize the initial distribution in the reactant state region.

The PMF relates the probability of sampling along the RC to a *free energy* profile. Higher probabilities are associated with relatively low free energy values compared to regions of lower probabilities. At the bottleneck (i.e., at $\min[\langle \tilde{\rho}(\xi) \rangle]$), the PMF corresponds to the barrier for a process under observation. The PMF, therefore, provides an intuitive *free energy surface* perspective to chemical dynamics. In terms of $w(\xi)$, the TST normalized density factor can be rewritten as

$$\langle \rho(\xi) \rangle = \frac{\exp[-\beta w(\xi)]}{\int_{-\infty}^{\infty} d\xi' \exp[-\beta w(\xi')]} \quad (29)$$

where the limits of integration in the denominator are over the reactant state region.

Another experimental measure of reactive chemical dynamics, capable of providing insights into nuclear quantum effects, is the kinetic isotope effect (KIE). The KIE is defined as the ratio of the reaction rate, k , involving the system with a lighter isotope (l) compared to the rate involving the system with a heavier (h) isotopic substitution

$$\text{KIE} = \frac{k^{(l)}}{k^{(h)}} \quad (30)$$

When the isotopic substitution involves a chemical bond pertaining to the rate-limiting step, the KIE is characterized as *primary*. If the substitution does not directly involve chemical bonds that are broken and formed, the KIE is referred to as *secondary*. This paper only addresses the

primary KIE for the intramolecular proton transfer reaction in malonaldehyde; however, the methodology presented here can be applied to estimating the secondary KIE as well.

As described in the previous section, the inclusion of zero-point motion and nuclear tunneling are important for calculating the rate of proton transfer. These nuclear quantum effects can be taken into account by using a *quantum* analog of TST based on the path integral formalism⁷²

$$k_{\text{PI-QTST}} = 1/2 \langle |d\xi_c/dt| \rangle_{\xi_c^{\ddagger}} \langle \rho(\xi_c^{\ddagger}) \rangle \quad (31)$$

where $d\xi_c/dt$ is the dynamical frequency factor and ρ is the density of sampled *centroid* RC values, defined as

$$\langle \rho(\xi_c) \rangle = \frac{\int d\mathbf{q}^{(1)} d\mathbf{q}^{(2)} \dots d\mathbf{q}^{(P)} \exp[-\beta \Phi(\mathbf{q}^{(1)}, \dots, \mathbf{q}^{(P)})] \delta[\tilde{\xi}_c(\mathbf{q}^{(c)}) - \xi_c]}{\int d\mathbf{q}^{(1)} d\mathbf{q}^{(2)} \dots d\mathbf{q}^{(P)} \exp[-\beta \Phi(\mathbf{q}^{(1)}, \dots, \mathbf{q}^{(P)})] h[\xi_c^{\ddagger} - \tilde{\xi}_c(\mathbf{q}^{(c)})]} \quad (32)$$

where $\beta = 1/k_B T$, h is the Heaviside step function, ξ_c^{\ddagger} is the location of the *dividing surface* partitioning the reactant and product regions, and $\tilde{\xi}_c$ is the value of the RC (as a function of the centroid coordinates $\mathbf{q}^{(c)} = 1/P \sum_{s=1}^P \mathbf{q}^{(s)}$) that was sampled during MD described by the effective potential $\Phi(\mathbf{q}^{(1)}, \dots, \mathbf{q}^{(P)})$ defined in eq 21. Similar to classical TST, the dynamical frequency factor can be estimated by the velocity of a free particle along the centroid RC direction

$$\langle |d\xi_c/dt| \rangle_{\xi_c^{\ddagger}} = \left(\frac{2}{\pi\beta} \right)^{1/2} \left\langle \left[\sum_{i=1}^{3N} \frac{1}{m_i} \left(\frac{\partial \xi_c(\mathbf{q}^{(c)})}{\partial q_i^{(c)}} \right)^2 \right]^{1/2} \right\rangle_{\xi_c^{\ddagger}} \quad (33)$$

Furthermore, the centroid density can be recast in terms of a *quantum* PMF $w(\xi_c)$ as

$$\langle \rho(\xi_c) \rangle = \frac{\exp[-\beta w(\xi_c)]}{\int_{-\infty}^{\infty} d\xi'_c \exp[-\beta w(\xi'_c)]} \quad (34)$$

where the integration limits are again over the RS region.

From eq 31, it is possible to compute the KIE for proton transfer in malonaldehyde by directly forming the ratio of the rates of reaction, i.e., $\text{KIE} = k_{\text{PI-QTST}}^{(\text{H})}/k_{\text{PI-QTST}}^{(\text{D})}$, where (H) denotes the system with the hydrogen isotope and (D) indicates the deuterium isotopic substitution. Alternatively, one can estimate the ratio of the hydrogen and deuterium centroid densities using thermodynamic integration (TI) with respect to mass^{73,74}

$$\frac{\langle \rho(\xi_c^{\ddagger}) \rangle^{(\text{H})}}{\langle \rho(\xi_c^{\ddagger}) \rangle^{(\text{D})}} = \exp \left\{ - \int_0^1 d\lambda \frac{\partial}{\partial \lambda} \ln[\rho(\lambda)] \right\} \quad (35)$$

where the parameter λ interpolates between the masses of the hydrogen and deuterium isotopes as

$$\tilde{m}_i(\lambda) = (1 - \lambda)m_i^{(\text{H})} + \lambda m_i^{(\text{D})} \quad (36)$$

Substituting the centroid density expression [eq 32] into the logarithmic derivative, we obtain

$$\frac{\langle \rho(\xi_c^z) \rangle^{(H)}}{\langle \rho(\xi_c^z) \rangle^{(D)}} = \exp \left[-\beta \int_0^1 d\lambda \left(\left\langle \frac{d\Phi(\lambda)}{d\lambda} \right\rangle_{RS} - \left\langle \frac{d\Phi(\lambda)}{d\lambda} \right\rangle_{\xi_c^z} \right) \right] \quad (37)$$

where $\langle \dots \rangle_{RS}$ denotes the canonical average over the reactant state region, $\langle \dots \rangle_{\xi_c^z}$ denotes the conditional average with the system constrained to the dividing surface ξ_c^z , and $d\Phi/d\lambda$ can be computed from the PI virial-like estimator^{65,75}

$$\frac{d\Phi(\lambda)}{d\lambda} \simeq - \sum_{i=1}^N \frac{d\tilde{m}_i/d\lambda}{\tilde{m}_i} \left[\frac{3}{2\beta} + \frac{1}{2P} \left(\sum_{s=1}^P (\mathbf{q}_i^{(s)} - \mathbf{q}_i^{(c)}) \cdot \frac{\partial V(\mathbf{q}^{(s)})}{\partial \mathbf{q}_i^{(s)}} \right) \right] \quad (38)$$

All elements in the above equation are previously defined. While two separate trajectories are required for each value of λ , the savings in computation (compared to explicit construction of the PMFs) can be substantial if one needs to perform multiple umbrella sampling trajectories in order to map out the entire RC range due to an intrinsic high free energy barrier. $d\Phi/d\lambda$ is typically a smooth, slowly varying function and thus can be numerically integrated using the simple trapezoidal rule along uniformly spaced λ points. Because the isotopic ratio of the dynamical frequency factors only involves conditional averages with the system constrained to ξ_c^z , the computation is not expensive, and thus it is not necessary to consider TI for this factor.

An alternative to PI-QTST for incorporating nuclear quantum effects in thermal rate calculations is the quantum instanton (QI) approach.^{32–35} Here, we only briefly sketch the key equations for QI theory and the steps for computing the relevant quantities using PIMD. The derivation of the QI rate equation begins with the formally exact quantum mechanical expression for the thermal rate constant:³²

$$k(T) Q_r(T) \equiv kQ_r = \frac{1}{2\pi\hbar} \int dE \exp(-\beta E) \times \frac{(2\pi\hbar)^2}{2} \text{tr}[\hat{F}_a \delta(E - \hat{H}) \hat{F}_b \delta(E - \hat{H})] \quad (39)$$

where $Q_r(T)$ is the reactant partition function per unit volume at temperature T , β is the inverse temperature $1/k_B T$, and $\hat{F}_\gamma = i\hbar[\hat{H}, h(\xi_\gamma(\mathbf{q}))]$ is the flux operator where the Heaviside step function h defines the location of the dividing surface $\xi_\gamma(\mathbf{q}) = 0$. Both the microcanonical density operator $\delta(E - \hat{H})$ and the integral over E can be evaluated within the steepest descent approximation to give the following approximate expression for the QI rate constant:³⁴

$$k(T) \simeq k_{QI} = \frac{1}{Q_r} C_{ff}(0) \frac{\sqrt{\pi} \hbar}{2 \Delta H} \quad (40)$$

In the above equation, $C_{ff}(0)$ is the zero time value of the flux–flux correlation function generalized to the case of two separate dividing surfaces³⁵

$$C_{ff}(t) = \text{tr}[e^{-\beta\hat{H}/2} \hat{F}_a e^{-\beta\hat{H}/2} e^{i\hat{H}t/\hbar} \hat{F}_b e^{-i\hat{H}t/\hbar}] \quad (41)$$

and ΔH is a particular type of energy variance

$$\Delta H^2 = \frac{\text{tr}[\hat{\Delta}_a e^{-\beta\hat{H}/2} \hat{H}^2 \hat{\Delta}_b e^{-\beta\hat{H}/2}] - \text{tr}[\hat{\Delta}_a e^{-\beta\hat{H}/2} \hat{H} \hat{\Delta}_b e^{-\beta\hat{H}/2} \hat{H}]}{\text{tr}[\hat{\Delta}_a e^{-\beta\hat{H}/2} \hat{\Delta}_b e^{-\beta\hat{H}/2}]} \quad (42)$$

with $\hat{\Delta}_a$ and $\hat{\Delta}_b$ being a modified version of the Dirac Δ function:

$$\hat{\Delta}_\gamma \equiv \Delta[\xi(\mathbf{q}) - \xi_\gamma] = \sqrt{\sum_{i=1}^N m_i^{-1} (\nabla_i \xi_\gamma(\mathbf{q}))^2} \times \delta[\xi(\mathbf{q}) - \xi_\gamma] \quad (43)$$

Rewriting eq 40 in the form

$$k_{QI} = \frac{C_{dd}(0)}{Q_r} \left\{ \frac{C_{ff}(0)}{C_{dd}(0)} \frac{\sqrt{\pi} \hbar}{2 \Delta H} \right\} \quad (44)$$

leads to components that are easily computable using PIMD, where

$$\begin{aligned} & \frac{C_{dd}(0; \xi_a, \xi_b)}{Q_r} \\ & \int d\mathbf{q}^{(1)} d\mathbf{q}^{(2)} \dots d\mathbf{q}^{(P)} \exp[-\beta\Phi(\{\mathbf{q}^{(s)}\})] \Delta \times \\ & \frac{[\xi(\mathbf{q}^{(P)}) - \xi_a] \Delta[\xi(\mathbf{q}^{(P/2)}) - \xi_b]}{\int d\mathbf{q}^{(1)} d\mathbf{q}^{(2)} \dots d\mathbf{q}^{(P)} \exp[-\beta\Phi(\{\mathbf{q}^{(s)}\})] \times} \\ & h[\xi^z - \xi(\mathbf{q}^{(P)})] h[\xi^z - \xi(\mathbf{q}^{(P/2)})] \end{aligned} \quad (45)$$

$$C_{ff}(0)/C_{dd}(0) = \langle f_v(\{\mathbf{q}^{(s)}\}) \rangle_{\xi(\mathbf{p}), \xi(\mathbf{p}/2)} \quad (46)$$

$$\Delta H^2 = \frac{1}{2} \langle F(\{\mathbf{q}^{(s)}\})^2 + G(\{\mathbf{q}^{(s)}\}) \rangle_{\xi(\mathbf{p}), \xi(\mathbf{p}/2)} \quad (47)$$

The conditional average $\langle \dots \rangle_{\xi(\mathbf{p}), \xi(\mathbf{p}/2)}$ is computed from the ensemble sampled with the P and $P/2$ PI slices constrained to the dividing surfaces

$$\begin{aligned} & \langle \dots \rangle_{\xi(\mathbf{p}), \xi(\mathbf{p}/2)} \\ & \int d\mathbf{q}^{(1)} d\mathbf{q}^{(2)} \dots d\mathbf{q}^{(P)} \exp[-\beta\Phi(\{\mathbf{q}^{(s)}\})] \Delta \times \\ & \frac{[\xi(\mathbf{q}^{(P)}) - \xi_a] \Delta[\xi(\mathbf{q}^{(P/2)}) - \xi_b] \times (\dots)}{\int d\mathbf{q}^{(1)} d\mathbf{q}^{(2)} \dots d\mathbf{q}^{(P)} \exp[-\beta\Phi(\{\mathbf{q}^{(s)}\})] \Delta \times} \\ & [\xi(\mathbf{q}^{(P)}) - \xi_a] \Delta[\xi(\mathbf{q}^{(P/2)}) - \xi_b] \end{aligned} \quad (48)$$

where the quantities within the average are defined as follows:

$$\begin{aligned} f_v(\{\mathbf{q}^{(s)}\}) &= \left(\frac{iP}{2\hbar\beta} \right)^2 \frac{\sum_{i=1}^N \nabla_i \xi_a(\mathbf{q}^{(P)}) \cdot (\mathbf{q}_i^{(1)} - \mathbf{q}_i^{(P-1)})}{\sqrt{\sum_{i=1}^N m_i^{-1} [\nabla_i \xi_a(\mathbf{q}^{(P)})]^2}} \times \\ & \frac{\sum_{i=1}^N \nabla_i \xi_b(\mathbf{q}^{(P/2)}) \cdot (\mathbf{q}_i^{(P/2+1)} - \mathbf{q}_i^{(P/2-1)})}{\sqrt{\sum_{i=1}^N m_i^{-1} [\nabla_i \xi_b(\mathbf{q}^{(P/2)})]^2}} \end{aligned} \quad (49)$$

$$F(\{\mathbf{q}^{(s)}\}) = -\frac{P}{\hbar^2\beta^2} \left\{ \sum_{s=1}^{P/2} - \sum_{s=P/2+1}^P \right\} \sum_{i=1}^N m_i (\mathbf{q}_i^{(s)} - \mathbf{q}_i^{(s-1)})^2 + \frac{2}{P} \left\{ \sum_{s=1}^{P/2-1} - \sum_{s=P/2+1}^{P-1} \right\} V(\mathbf{q}^{(s)}) \quad (50)$$

$$G(\{\mathbf{q}^{(s)}\}) = \frac{2fP}{\beta^2} - \frac{4P}{\hbar^2\beta^3} \sum_{s=1}^P \sum_{i=1}^N m_i (\mathbf{q}_i^{(s)} - \mathbf{q}_i^{(s-1)})^2 \quad (51)$$

In the expression for G , f is the total number of degrees of freedom (i.e., $f = 3N$ in our malonaldehyde application).

All factors required to calculate the QI rate can be computed using the PIMD facilities within Amber.²⁸ For example, the joint distribution [eq 45] can be computed using umbrella sampling along the reaction coordinates of the P and $P/2$ slices. A set of two-dimensional (2D) biased simulations, each enhancing the sampling near a particular point of the 2D ($\xi_P \times \xi_{P/2}$) configurational space, is required to map out the entire QI joint distribution. Using the weighted histogram analysis method (WHAM),^{71,76,77} the generated biased distributions can be unbiased to form $C_{dd}(0)/Q_r$ on the DG-EVB ground-state surface. The remaining factors, involving the conditional average of f_v , F , and G , can be computed again using umbrella sampling with the P and $P/2$ slices constrained to the dividing surface ξ^z . Now, one can estimate KIEs from QI calculations by directly forming the ratio of the rates or by TI integration with respect to mass. For the latter case, the canonical average is defined over the 2D configurational space of the RS region delineated by the two dividing surfaces, and the conditional average is computed with the P and $P/2$ slices constrained to the dividing surface ξ^z . Also, the modified Dirac Δ functions in the conditional average of eq 48 introduce an additional term in the estimator for the logarithmic derivative of the delta-delta correlation function

$$-\beta^{-1} \frac{d \ln C_{dd}(\lambda)}{d\lambda} \approx -\sum_{i=1}^N \frac{d\tilde{m}_i/d\lambda}{\tilde{m}_i} \left[\frac{3}{2\beta} + \frac{1}{2P} \left(\sum_{s=1}^P (\mathbf{q}_i^{(s)} - \mathbf{q}_i^{(c)}) \cdot \frac{\partial V(\mathbf{q}^{(s)})}{\partial \mathbf{q}_i^{(s)}} \right) - \frac{\sum_{\gamma=a,b} |\nabla_{i,\xi_\gamma}(\mathbf{q})|^2}{2\tilde{m}_i \left\| \sum_{i=1}^N \tilde{m}_i^{-1} (\nabla_{i,\xi_\gamma}(\mathbf{q})) \right\|^2} \right] \quad (52)$$

Thus, whereas TI integration with respect to mass within PIMD uses the same estimator for sampling at the dividing surface and in the RS, QI uses eq 52 for the sampling at the dividing surfaces and eq 38 for sampling in the RS, the difference being the last term in eq 52.

3. Simulation Details

The symmetric intramolecular proton transfer reaction in malonaldehyde is depicted in Figure 1. All classical and quantum nuclear molecular dynamics simulations were performed using the DG-EVB facility within Amber 11.²⁸ The DG-EVB surface was fit using the energy, gradient, and

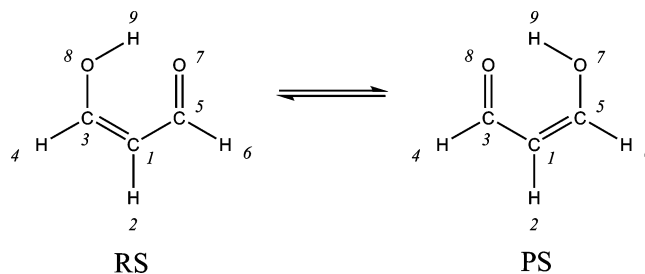


Figure 1. Intramolecular proton transfer reaction in malonaldehyde.

Hessian information at the RS, TS, PS, and off-TS geometries. The geometries at the RS, TS, and PS were optimized according to the WIBD^{78,79} model chemistry using a development version of the Gaussian suite.⁸⁰ The additional nonoptimized off-TS geometry lies *in front* of the TS and corresponds to both OH bonds being 1.5 Å. This point was chosen to improve the surface fit to the repulsive wall in front of the TS. Being able to systematically incorporate additional ab initio points as needed for refining the surface fit is one strength of the DG-EVB approach. The gradients and Hessian tensors were computed from the resulting geometries at the CCSD/cc-pVTZ level of theory. This choice of ab initio methods is comparable to the most accurate approach described in the literature for malonaldehyde, employing frozen-core CCSD(T)/aug-cc-pV5Z single-point calculations at frozen-core CCSD(T)/aug-cc-pVTZ optimized C_s and C_{2v} structures.³⁶ The present calculations yield a barrier height of 4.08 kcal/mol, nearly identical to the barrier of 4.09 kcal/mol obtained by Bowman.

The diabatic state Hamiltonian matrix element is approximated as a harmonic expansion about the ab initio optimized minimum (RS or PS) plus a scaled nonbonding, van der Waals, exponential-6 term from the universal force field (UFF)⁸¹ and Amber force field terms for angles and dihedrals:

$$H_{NN}(\mathbf{q}) = C + \mathbf{G}_N \cdot \Delta \mathbf{q}_r + \frac{\Delta \mathbf{q}_r^i \cdot \tilde{\mathbf{K}}_N \cdot \Delta \mathbf{q}_r}{2} + \eta \sum_i A_{\text{UFF}}^i \exp[-B_{\text{UFF}}^i \Delta \mathbf{q}_r^i] + \sum_{\text{angles}} K_\theta (\mathbf{q}_\theta - \theta_0)^2 + \sum_{\text{dihedrals}} (V_n/2)(1 + \cos[n\mathbf{q}_\phi - \delta]) \quad (53)$$

In the above equation, A_{UFF}^i and B_{UFF}^i are UFF exponential-6 parameters, $\Delta \mathbf{q}_r^i$ is the selected repulsive coordinate (e.g., the distance between the transferring H and the acceptor O in malonaldehyde) for H_{NN} and

$$C = \varepsilon_\psi(\mathbf{q}_N) - \eta \sum_i A_{\text{UFF}}^i \exp[-B_{\text{UFF}}^i \Delta \mathbf{q}_r^i] - \sum_{\text{angles}} K_\theta (\mathbf{q}_\theta - \theta_0)^2 - \sum_{\text{dihedrals}} (V_n/2)(1 + \cos[n\mathbf{q}_\phi - \delta]) \quad (54)$$

is a constant ensuring that the ab initio energy is recovered at the DG-EVB data point, $\mathbf{q} = \mathbf{q}_N$. A set of redundant internal coordinates, $\mathbf{q} = \{\mathbf{q}_r, \mathbf{q}_\theta, \mathbf{q}_\phi\}$, comprised of bond lengths (r), angles (θ), and dihedrals (ϕ) is used in the DG-EVB method to maintain invariance of the resulting energy hypersurface under global rotations. Typically, $H_{12}^2(\mathbf{q})$ does

not depend on the full set of redundant internal coordinates, and a subspace comprised of coordinates that change within a prescribed tolerance between the reactant and product configurations has been shown to provide sufficient accuracy.²⁶ For the malonaldehyde system, the above parameters for the angle (K_θ, θ_0) and dihedral (V_n, n, δ) interactions are taken from GAFF (generalized Amber force field).⁶ The UFF interactions were scaled by $\eta = 0.60$, and an *optimized* average value of 0.85 was used for all Gaussian α_K parameters in this study.²⁶

The *classical* RC ξ , chosen to represent the breaking and formation of a chemical bond, is defined as the difference of bond lengths r between the donor (D), the acceptor (A), and the transferring particle (H) positions

$$\xi(\mathbf{q}) = r(\mathbf{q}_D, \mathbf{q}_H) - r(\mathbf{q}_A, \mathbf{q}_H) \quad (55)$$

$\xi(\mathbf{q}) < 0$, thus, represents the reactant state region, while $\xi(\mathbf{q}) > 0$ represents the product state and $\xi(\mathbf{q}) = 0$ delineates the transition state (TS). Within the umbrella sampling framework,⁷¹ the system Hamiltonian is described by the modified potential

$$\begin{aligned} V_{\text{biased}}^{(n)}(\mathbf{q}) &= \varepsilon_0(\mathbf{q}) + V_{\text{umb}}^{(n)}(\mathbf{q}) \\ &= \varepsilon_0(\mathbf{q}) + \frac{1}{2}k^{(n)}[\xi(\mathbf{q}) - \xi_0^{(n)}(\mathbf{q})]^2 \end{aligned} \quad (56)$$

where \mathbf{q} is the set of system coordinates, k is the harmonic force constant parameter, and $V_{\text{umb}}^{(n)}$ is a biasing potential that is added to the original system potential ε_0 (obtained from diagonalization of the EVB matrix) to enhance the sampling of a predetermined region of configuration space near $\xi_0^{(n)}$. The distributions of $\xi^{(n)}$ from all of the (n) biased simulations are unbiased and combined using WHAM^{71,76} to form the PMF describing the chemical reaction of interest. All biased sampling simulations were performed in the NVT ensemble at a temperature of 300 K using a leapfrog Verlet integration time step of 0.5 fs for a minimum of 3 ns of total sampling per window.⁸² The temperature was maintained via coupling of the system to a Langevin thermostat⁸³ with a collision frequency of 1 ps^{-1} .

With the exception of the modifications described below, similar simulation protocols were followed for the PIMD simulations. The RC chosen to describe the breaking and formation of a chemical bond within the PI-QTST framework is the difference of bond lengths r between the donor, the acceptor, and transferring particle centroid coordinates

$$\tilde{\xi}_c(\mathbf{q}) = r(\mathbf{q}_D^{(c)}, \mathbf{q}_H^{(c)}) - r(\mathbf{q}_A^{(c)}, \mathbf{q}_H^{(c)}) \quad (57)$$

where the centroid position for particle χ is defined as the average of positions over the P path integral slices

$$\mathbf{q}_\chi^{(c)} = \frac{1}{P} \sum_{s=1}^P \mathbf{q}_\chi^{(s)} \quad (58)$$

Substituting $\tilde{\xi}_c$ for ξ in eq 56 allows for enhanced sampling of a prescribed reaction path using the same umbrella sampling procedure as employed for the classical molecular dynamics. The *quantum* PMF then is generated using WHAM^{71,76} from the centroid densities sampled in all the

biased trajectories. In contrast to the classical simulations described above, the PIMD was propagated in the normal mode representation⁶⁶ using a 0.5 fs leapfrog Verlet integration time step and with the temperature maintained at 300 K via coupling to a Nosé–Hoover chain bath of size four. Each region of the reaction path in the vicinity of $\xi_0^{(n)}$ is sampled for a minimum of 3 ns.

Two procedures were utilized for estimating the primary KIE in malonaldehyde. In the direct approach, the PI-QTST⁷² rates for the intramolecular proton transfer of the hydrogen and deuterium isotopes are computed separately using eq 31 to form the ratio $k_{\text{PI-QTST}}^{(\text{H})}/k_{\text{PI-QTST}}^{(\text{D})}$. The centroid density component of the PI-QTST rate was computed using umbrella sampling with $k^{(n)} = 100.0 \text{ kcal/mol } \text{Å}^2$ and $\xi_0^{(n)} = \{-0.60, -0.40, -0.20, 0.0, 0.20, 0.40, 0.60\} \text{ Å}$. The dynamical frequency factor was computed again with umbrella sampling with the RC restrained to the transition state region, i.e., $\xi_0 = 0.0 \text{ Å}$, using a harmonic force constant of $k = 2000.0 \text{ kcal/mol } \text{Å}^2$. A total of 14 independent biased trajectories (seven for each isotopically labeled proton transfer) are required for estimating the ratio of centroid densities, while two independent biased simulations are required for estimating the ratio of the frequency factors. In the TI with respect to mass approach,^{73,74} the ratio of isotopic centroid densities is estimated using eq 37 from a set of $\lambda = \{0.0, 0.2, 0.4, 0.6, 0.8, 1.0\}$ simulations where the average $\langle \dots \rangle_{\text{RS}}$ is computed from ground-state EVB sampling in the RS region and the average $\langle \dots \rangle_{\xi_c^{\neq}}$ is computed from umbrella sampling with the RC restrained to $\xi_0 = 0.0 \text{ Å}$ via a harmonic force constant, $k = 2000.0 \text{ kcal/mol } \text{Å}^2$. The above TI by mass, therefore, entails a total of $6\lambda \times 2$ trajectories/ $\lambda = 12$ independent trajectories. Although the computational costs appear similar for both alternatives, it is expected that the statistical error for the KIE estimate based on relative rates [from eq 35] will be smaller compared to the estimate based on absolute rates [from eq 31]. For cases where a large number of biased trajectories are required to cover the range of RC values, due to an intrinsically high free energy barrier of the system or because of multiple dividing surfaces, the TI by mass route will be computationally less expensive.

Particularly in the case of QI^{33,34} calculations, TI by mass^{73,74} is a substantially more efficient approach for estimating KIEs. In the direct approach, the conditional averages of f_v , F , and G [eqs 49–51] were computed using umbrella sampling with the P and $P/2$ PI slices restrained to the dividing surface with a harmonic force constant of $k = 2000.0 \text{ kcal/mol } \text{Å}^2$. The 2D joint distribution was computed using umbrella sampling with $k^{(n)} = 100.0 \text{ kcal/mol } \text{Å}^2$ and $\xi_0^{(n)} = \{-1.00, -0.80, -0.60, -0.40, -0.20, 0.0, 0.20, 0.40, 0.60, 0.80, 1.00\} \text{ Å}$ for all combinations of pairs of restraints applied to the P and $P/2$ PI slices. Thus, we have a grid of $11 \times 11 = 121$ independent biased sampling trajectories for each isotope, for a total of 242 trajectories in a single KIE estimate. Compared to the 12 trajectories required for TI by mass, the computation disparity is significant. A protocol similar to the above PI-QTST details was used for the QI TI with respect to the mass procedure, with the only exception being that the conditional average in eq 37 is computed from umbrella sampling restraining the RC of both

Table 1. Computed TST, PI-QTST, and QI Chemical Rates and KIE for the Intramolecular Proton Transfer Reaction in Malonaldehyde^a

$$k_{\text{TST}} = \frac{1}{2} \langle \dot{\xi} \rangle_{\xi^{\ddagger}} \langle \rho(\xi^{\ddagger}) \rangle$$

isotope	$\langle \dot{\xi} \rangle_{\xi^{\ddagger}} [\text{\AA} \cdot \text{s}^{-1}]$	$\langle \rho(\xi^{\ddagger}) \rangle [\text{\AA}^{-1}]$	$k_{\text{TST}} [\text{s}^{-1}]$	KIE
H	$2.48 \times 10^{13} (\pm 1.31 \times 10^9)$	$2.04 \times 10^{-3} (\pm 1.81 \times 10^{-4})$	$2.53 \times 10^{10} (\pm 2.25 \times 10^9)$	1.54 (± 0.16)
D	$1.78 \times 10^{13} (\pm 1.18 \times 10^9)$	$1.84 \times 10^{-3} (\pm 1.04 \times 10^{-4})$	$1.64 \times 10^{10} (\pm 9.28 \times 10^8)$	

$$k_{\text{PI-QTST}} = \frac{1}{2} \langle |\dot{\xi}_c| \rangle_{\xi_c^{\ddagger}} \langle \rho(\xi_c^{\ddagger}) \rangle$$

isotope	$\langle \dot{\xi}_c \rangle_{\xi_c^{\ddagger}} [\text{\AA} \cdot \text{s}^{-1}]$	$\langle \rho(\xi_c^{\ddagger}) \rangle [\text{\AA}^{-1}]$	$k_{\text{PI-QTST}} [\text{s}^{-1}]$	KIE
H	$2.46 \times 10^{13} (\pm 4.53 \times 10^8)$	$5.34 \times 10^{-2} (\pm 2.77 \times 10^{-3})$	$6.56 \times 10^{11} (\pm 3.40 \times 10^{10})$	4.05 (± 0.27)
D	$1.77 \times 10^{13} (\pm 7.94 \times 10^8)$	$1.83 \times 10^{-2} (\pm 7.31 \times 10^{-4})$	$1.62 \times 10^{11} (\pm 6.47 \times 10^9)$	

$$k_{\text{QI}} = \frac{C_{dd}(0)}{Q_r} \left\{ \frac{C_{ff}(0)}{C_{dd}(0)} \frac{\sqrt{\pi} \hbar}{2 \Delta H} \right\}$$

isotope	$\left\{ \frac{C_{ff}(0)}{C_{dd}(0)} \frac{\sqrt{\pi} \hbar}{2 \Delta H} \right\} [\text{\AA}^2 \text{s}^{-1}]$	$\frac{C_{dd}(0)}{Q_r} [\text{\AA}^{-2}]$	$k_{\text{QI}} [\text{s}^{-1}]$	KIE
H	$3.71 \times 10^{12} (\pm 9.61 \times 10^{11})$	$5.98 \times 10^{-2} (\pm 9.76 \times 10^{-4})$	$2.22 \times 10^{11} (\pm 5.76 \times 10^{10})$	2.41 (± 0.86)
D	$3.44 \times 10^{12} (\pm 8.38 \times 10^{11})$	$2.67 \times 10^{-2} (\pm 6.02 \times 10^{-4})$	$9.19 \times 10^{10} (\pm 2.25 \times 10^{10})$	

^a The uncertainties given within parentheses are estimated from the distribution of results as a parameter of the MD sampling intervals.

the P and $P/2$ slices to $\xi_0 = 0.0 \text{ \AA}$ via a harmonic force constant of $k = 2000.0 \text{ kcal/mol \AA}^2$, i.e., $\langle \dots \rangle_{\xi_c^{\ddagger}} \rightarrow \langle \dots \rangle_{\xi_{(P)}}$, $\xi_{(P/2)}$. The relevant estimator used for the constrained sampling on the dividing surfaces is given by eq 52, while sampling within the RS well uses the estimator defined in eq 38.

Estimating the uncertainty in a rate calculation is not straightforward, especially when the quantities are derived from sampling over phase space. The frequency factor, PMF, and TI by mass calculations are subject to uncertainties related to the level of convergence of phase space sampling. To provide an estimate of this type of uncertainty, we compute the various components as a parameter of MD sampling intervals. The uncertainties reported in Table 1, using this approach, give an indication of the variability of the results as a parameter of phase space averaging. Adequate MD sampling should give averages with a relatively small standard deviation.

4. Results and Discussion

To evaluate the effectiveness of the DG-EVB method for constructing ab initio based reactive PESs for modeling chemical dynamics, the classical TST rate, PI-QTST rate, QI rate, and KIE for the prototypical intramolecular proton transfer reaction in malonaldehyde are computed. The calculations of these observables require MD sampling of configurational space to obtain thermodynamic averages. Furthermore, since the chemistry under consideration involves a proton, nuclear quantum effects, such as zero-point motion and tunneling, are important for describing the

reaction rate. These elements have been integrated into the latest release of the Amber biosimulation suite (version 11),²⁸ and the results for malonaldehyde are described below.

The 2D PES for the intramolecular proton transfer reaction in malonaldehyde, employing four Gaussian centers (located at the RS, TS, PS, and off-TS) to fit H_{12}^2 , is depicted in Figure 2. The diabatic state H_{NV} is represented as a quadratic expansion about the ab initio minimum configuration, augmented by a nonbonding, van der Waals, exponential-6 term from the universal force field⁸¹ and Amber angle and dihedral terms [eq 53]. The UFF term was included to prevent an anomalous “swimming hole” located behind the TS region geometry.²⁷ Alternatively, one can fill in this swimming hole by placing additional distributed Gaussians in these anomalous regions. The minima are indicated with a green dot and the TS with an orange dot. The additional configuration (indicated with a blue dot) along the plane orthogonal to the reaction path improves the curvature of the surface in this region. The goal here is to develop a robust, automated method capable of generating a multidimensional surface from ab initio information gleaned from sparse and strategically chosen geometries. For the malonaldehyde system, utilizing a quadratic expansion about the ab initio minimum and empirical FF terms is adequate for generating an accurate smooth PES for exploring chemical dynamics. The \mathbf{B} vector of eq 14 was solved using the GMRES^{57–60} algorithm with a convergence tolerance of 1×10^{-9} . The maximum error between the resulting DG-EVB and ab initio energies for geometries at the fit points was computed to be $7.49 \times 10^{-5} \text{ kcal/mol}$. The correspond-

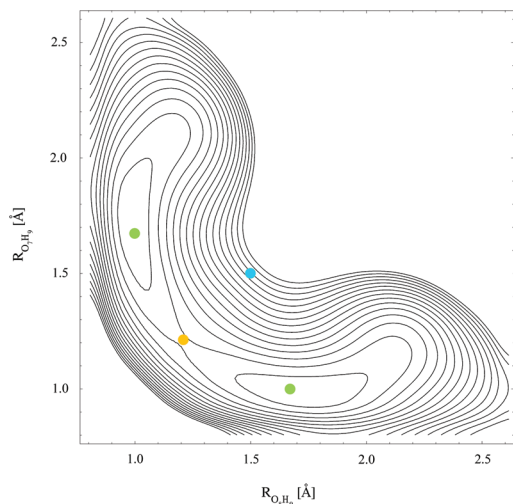


Figure 2. Two-dimensional potential energy surface for the intramolecular proton transfer reaction in malonaldehyde employing a four Gaussian fit for H_{12} and H_{MN} with a UFF repulsive term and Amber angle and dihedral terms. The minima are denoted with a green dot and the TS with an orange dot. An additional off-TS point (depicted by the blue dot) along the plane orthogonal to the reaction path improves the overall fit. A grid of geometries obtained from relaxed scans at the Hartree–Fock level and the cc-pVTZ basis set was used in the DG-EVB energy calculations.

ing maximum root-mean-square deviations between the DG-EVB and ab initio gradients is 5.65×10^{-4} kcal/mol Å, and the maximum difference between elements of the DG-EVB and ab initio internal coordinate Hessian matrices is 4.86×10^{-8} hartree/(internal coordinate)². Our DG-EVB surface exhibits an energy barrier height of 4.08 kcal/mol, which is in excellent agreement with the best estimate of 4.09 kcal/mol published in the literature.³⁶

With a well-defined PES, it is now possible to perform nuclear dynamics on this surface incorporating ab initio data. Since the ab initio energy barrier of 4.08 kcal/mol is well above thermal energy, conventional MD may not adequately sample important TS configurations. The resulting PMF (Figure 3) obtained from ground-state dynamics indicates that malonaldehyde predominantly samples the conformational space of the reactant and product region but not so much the barrier region. Figure 3 is obtained by initiating conventional MD on the reactant minimum and collecting the statistics (indicated by the gray histogram in the background) for sampling a particular value of the RC. At an average temperature of 300 K, the system has sufficient thermal energy to overcome the energy barrier. Sampling near the TS, however, remains statistically insignificant. This intrinsic *rare event* nature of the chemistry, thus, requires enhanced sampling techniques for sampling conformations within the important TS region.

Utilizing the umbrella sampling procedure⁷¹ described in the methodology section, a set of independent biased trajectories was used to map out the distribution of RC values over the entire range of the reactive path. Each trajectory only enhances the sampling about a predefined RC value, $\xi_0^{(n)}$. Using the WHAM^{71,76} procedure, the set of biased distributions (gray curves, Figure 4) are combined to form a

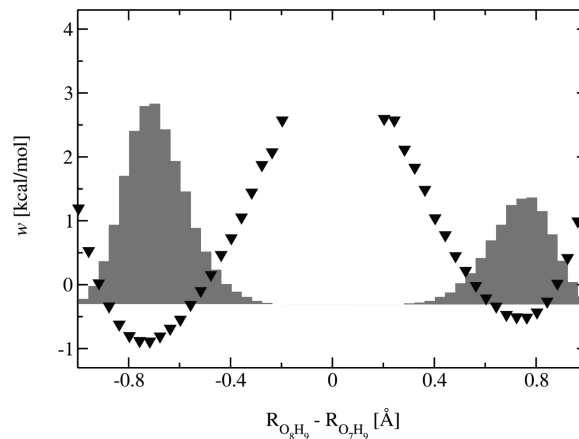


Figure 3. Potential of mean force obtained from direct sampling on the DG-EVB ground-state surface (▼). The gray background depicts the histogram of RC values encountered during the MD.

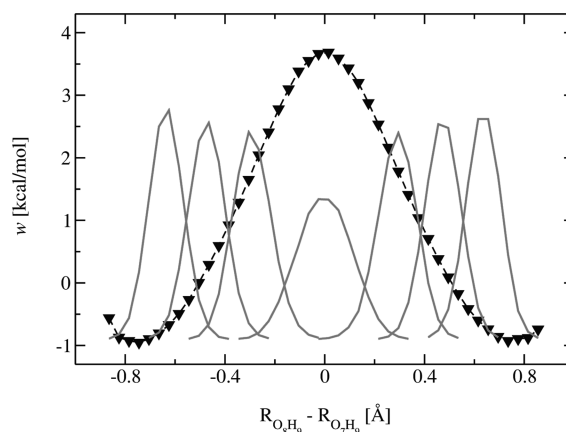


Figure 4. Potential of mean force obtained from umbrella sampling on the DG-EVB ground-state surface. The dashed line is a cubic spline fit through the data points (▼). The gray curves in the background show the histogram of RC values sampled along the biased MD trajectories, with $\xi_0^{(n)} = \{-0.6, -0.4, -0.2, 0.0, 0.2, 0.4, 0.6\}$ Å. These distributions are combined using WHAM.

free energy profile (▼curve, Figure 4) spanning the entire range of the RC for the proton transfer under observation. The resulting normalized density contribution [eq 29] to the TST rate is $2.04 \times 10^{-3} (\pm 1.81 \times 10^{-4})$ Å⁻¹. Again, using umbrella sampling with the RC restrained to the dividing surface ($\xi_0 = 0.0$ Å), the other contribution from the dynamical frequency factor can be estimated from the gradient of the RC. With a value of $2.48 \times 10^{13} (\pm 1.31 \times 10^9)$ Å s⁻¹ for $\langle d\xi/dt \rangle_{\xi^{\ddagger}}$, the classical TST estimate of the proton transfer rate is $2.53 \times 10^{10} (\pm 2.25 \times 10^9)$ s⁻¹.

While the above classical model of the proton experiences a free energy barrier of ~ 3.75 kcal/mol, zero-point motion and nuclear tunneling may effectively lower this reaction barrier. Figure 5 compares the quantum PMF (▼) obtained from PIMD sampling on the DG-EVB surface to the classical PMF (dotted line) from Figure 4. Both curves are normalized to their respective RS partition function. The free energy values at the TS ($w(\xi^{\ddagger})$ and $w(\xi_c^{\ddagger})$), therefore, are relative to absolute energy origins as defined in the classical TST

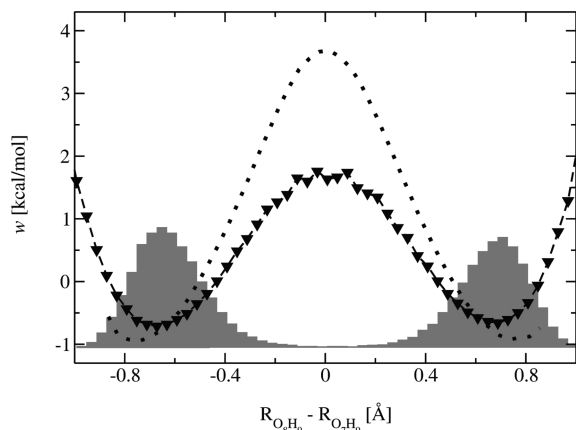


Figure 5. Potential of mean force obtained from direct PIMD sampling on the DG-EVB ground-state surface (▼). The gray background depicts the histogram of RC values encountered during the PIMD. For comparison, the classical PMF from Figure 4 is shown by the dotted curve.

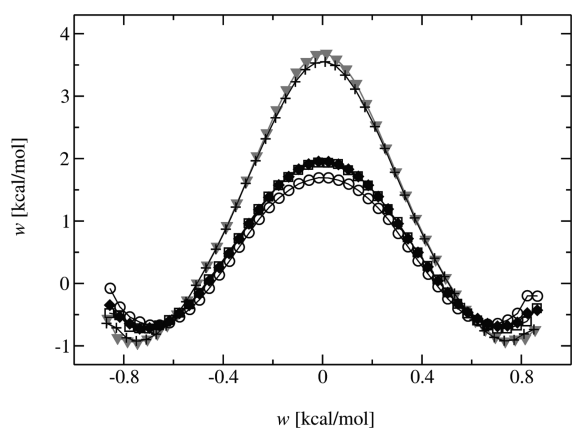


Figure 6. Potential of mean force as a parameter of the level of nuclear quantization: classical nuclei (▼), quantization only of the donor and acceptor oxygens (+), quantization of transferring proton (◆), quantization of donor and acceptor oxygens and transferring proton (□), full nuclear quantization (○).

and PI-QTST rate equations. It is seen that full nuclear quantization of malonaldehyde lowers the free energy barrier by about 2 kcal/mol. This result is similar to trends observed in earlier studies of malonaldehyde by Tuckerman and Marx⁸⁴ and by Schofield and Iftimie.^{85,86} While inclusion of nuclear quantization sufficiently lowers the free energy barrier to allow sampling of the TS, a general framework for enhancing sampling of a particular region of configurational space is desirable. Reactions in enzyme environments, for example, may encounter barriers typically on the order of ~ 10 kcal/mol or higher. Our implementation of umbrella sampling within the PIMD function in Amber will permit studies of rare event phenomena using quantum dynamics approaches.

From a computational efficiency perspective, it is also worthwhile to consider if nuclear quantization is necessary for the entire system or if certain degrees of freedom can remain in the classical description. Figure 6 shows the PMFs as a parameter of the *level of nuclear quantization*, where the ▼ curve reproduces the classical description from Figure

4 and the ○ curve depicts the fully quantum description. Here, all PMFs are obtained using the umbrella sampling protocol described in the Simulation Details section. The □ curve reflects quantization only of the donor, acceptor, and transferring proton, while the ◆ curve shows the impact of quantization only of the transferring proton. Both of these curves are on top of each other and are slightly higher than the fully quantized system. The remaining curve (+) corresponds to quantization of only the donor and acceptor oxygens. As expected from theoretical considerations, nuclear quantization of the light particles (hydrogens) has the most significant effect on the free energy barrier height; whereas, quantization of the heavier oxygen atoms only has a minor effect compared to the classical description. For the case of malonaldehyde, quantizing the transferring proton is sufficient to capture the bulk of the quantum effects associated with our definition of the RC. This trend may not be generally transferrable to other more flexible molecules or to alternative prescriptions of the reaction path. In addition to lowering the reaction barrier, nuclear quantization also shifts the location of the reactant and product minima and changes the curvature in the transition state region. This alteration of both the barrier height and shape of the free energy surface will manifest in KIE measurements.

The estimates of primary KIEs in this study were obtained using both the direct approach of explicitly forming the ratio of the absolute rates and via TI with respect to mass. Table 1 shows all contributions to the chemical rates for the hydrogen and deuterium isotopes as computed from classical TST, PI-QTST, and QI prescriptions. Because experimental measurements of KIEs for malonaldehyde are unavailable, the latter two approximate quantum rates serve to provide ballpark estimates from well-established methods. The KIE of $4.05 (\pm 0.27)$ from PI-QTST is about one and a half times larger than the QI value of $2.41 (\pm 0.86)$. The absolute rates from PI-QTST, however, are approximately two to three times larger than the estimates from QI, due predominantly to the larger frequency factors. The contribution to the rate from the PMF is smaller in PI-QTST than in QI for both proton and deuterium transfers. Since PI-QTS and QI are approximate quantum rate methods, equivalently at the level of transition state theory, the differences in the rates are not due to dynamical recrossing effects. A clear relationship between PI-QTS and the QI method has yet to be established. Furthermore, the results indicate that classical TST is inadequate in providing KIE estimates. Here, isotopic substitution predominantly impacts the frequency factor and leaves the PMF contribution to the rate little changed, resulting in a KIE of 1.54. The effective lowering of the barrier height due to zero-point motion and nuclear tunneling are missing. Nuclear quantization via PI recovers these effects and provides the dominant contributing factor of $2.91 (\pm 0.19)$ to the PI-QTST KIE. The frequency factors obtained from PIMD are similar to those computed from classical sampling, suggesting that the centroid coordinates of the system at the dividing surface for this symmetric transfer can be represented by classical nuclear coordinates. For asymmetric reactions, this correspondence between the centroid and classical variables may not hold, and one cannot

simply assume that the computationally less demanding classical MD sampling will be sufficient for estimating the frequency factor component of the PI-QTST rate.

While a direct comparison of our computed KIEs with the published results of Schofield⁸⁵ is complicated by differences in the underlying PESs, it is worthwhile to identify the contributing factors giving rise to varying estimates of the KIE. In their previous PI-QTST study employing a molecular mechanical EVB potential for describing the proton transfer reaction,⁸⁵ the primary KIEs range from 6.49 to 11.41 and depend on the choice of RC. The sensitivity of KIEs on the prescription of the RC is to be expected, as PI-QTST is a transition state approximation to the quantum rate. Different RCs may result in varying levels of barrier recrossings, and it is this recrossing factor that is missing in the KIE calculations. Furthermore, the molecular mechanical parameters employed in that EVB formulation give rise to an energy difference of 8.75 kcal/mol between the reactant and transition state conformations. This activation energy is about two times larger than the barrier obtained from the present DG-EVB approach, leading to rates that are about 2–3 orders of magnitude smaller. Also, a higher barrier typically leads to a larger KIE, and this trend is observed between the two methods.

Figure 7a shows the 2D PMF associated with the joint distribution of the RC along the P and $P/2$ PI slices for the proton transfer reaction. The contribution to the QI rate is obtained using the dividing surface corresponding to the top of the free energy barrier in Figure 7b with $\xi_a = \xi_b = \xi^{\ddagger} = 0.0$ Å. The deuterium transfer is depicted by ▼, and the proton transfer is depicted by ○. Note that this PMF is normalized with respect to the reactant partition function within the full 2D space, and as such, one cannot simply perform a 1D biased sampling with both dividing surfaces set equal to each other. The 2D distribution requires a total of $11 \times 11 = 121$ biased sampling trajectories in order to map out the entire $\xi_a \times \xi_b$ configurational space. For estimating the KIE, the computation becomes quite expensive, totaling 242 trajectories for the malonaldehyde system.

As noted in the methods section, the computationally expensive PMF calculation can be avoided if one reformulates the ratio of isotope densities as a thermodynamic integration over mass. Figure 8 displays the average values of $-\beta \partial \Phi(\lambda) / \partial \lambda$ sampled during PIMD in the RS well and at the dividing surface as a parameter of λ . Integration over the λ values using eq 37 provides a centroid density value of 3.08 (± 0.01), which is comparable to 2.91 (± 0.19) obtained directly from the PMFs. Similarly, TI by mass for the QI calculations provides an estimate of 4.73 (± 0.02) for $C_{dd}^{(H)}(0)/C_{dd}^{(D)}(0)$ compared to the value of 2.24 (± 0.06) from direct calculation. In the direct calculations, the QI PMF barrier height for proton transfer is 1.68 kcal/mol and is 2.16 kcal/mol for deuterium transfer [see Figure 7b]. To get a $C_{dd}^{(H)}(0)/C_{dd}^{(D)}(0)$ of ~ 4.73 , the proton transfer barrier height needs to be lowered to 1.23 kcal/mol (while keeping the deuterium transfer barrier at 2.16 kcal/mol) or the deuterium transfer barrier height needs to increase to 2.60 kcal/mol (while keeping the proton transfer barrier at 1.68 kcal/mol). This resolution difference of ~ 0.5 kcal/mol is especially

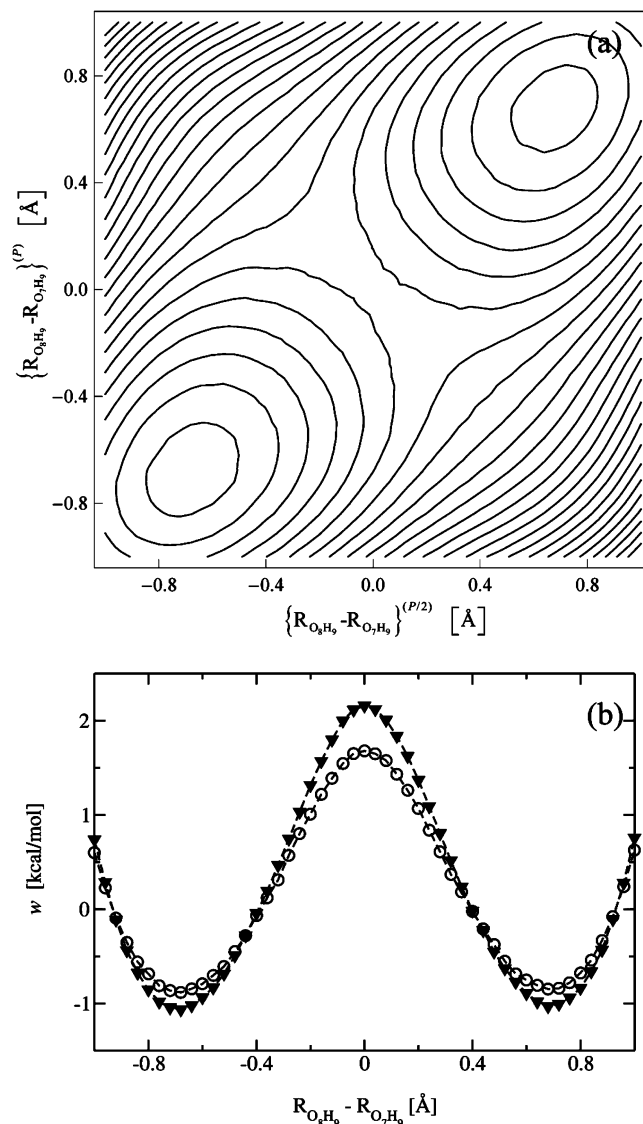


Figure 7. Potential of mean force associated with the joint probability density $C_{dd}(0; \xi_a, \xi_b) / Q_r$ of eq 45. (a) Contours for proton transfer (0.4 kcal/mol each) as a parameter of RC values along the P and $P/2$ PI slices. (b) Slice through the contours at $\xi_a = \xi_b$ that corresponds to a single dividing surface. The deuterium transfer is depicted by ▼, and the proton transfer is depicted by ○.

challenging to achieve in the QI 2D PMFs, where errors in the reweighting of the 121 biased distributions may accumulate and impact the global unbiased distribution. The virial estimator for the average quantities in TI by mass has been shown to converge efficiently.⁷⁴ The errors in the TI by mass estimate of the KIE, thus, are expected to be smaller than errors arising from the direct approach.

Multiplying the ratio of isotopic densities by the corresponding ratio of frequency factors, the PI-QTST estimate of 4.27 (± 0.01) for the KIE is within 80% of the QI value of 5.10 (± 1.81). The error in the QI estimate, however, is much larger than that from PI-QTST because of the variability of the frequency factor estimates. While both approaches average the frequency factor over the same time span (3 ns), QI appears to require more phase space sampling to achieve a higher level of convergence. On the other hand,

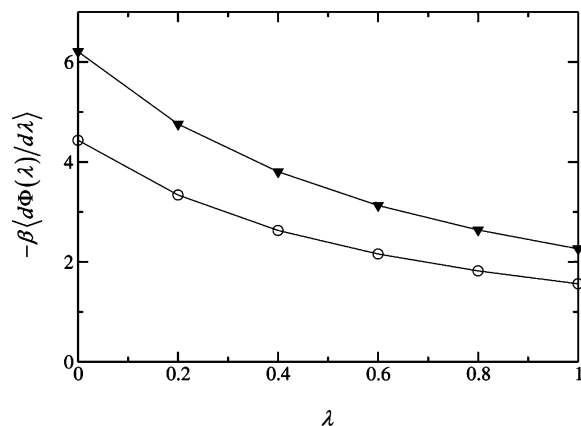


Figure 8. Average value of $-\beta\langle d\Phi(\lambda)/d\lambda\rangle$ sampled during PIMD in the RS region (\blacktriangledown) and at the dividing surface ξ_c^{\ddagger} (\circ) as a parameter of λ .

the phase space averages required for TI by mass converge much faster. A nanosecond of sampling of the virial estimators for $d\Phi/d\lambda$ at each value of λ is sufficient for obtaining the ratio of isotopic densities with relative uncertainties that are less than 1%. The savings in computation as well as a much smaller estimate of the uncertainty is especially evident in QI calculations, where the two sets of computationally demanding PIMD sampling can be avoided altogether by TI integration over the mass.

5. Concluding Remarks

This work presents an application of the distributed Gaussian EVB method for constructing an ab initio-based reactive potential energy surface describing the intramolecular proton transfer reaction in malonaldehyde. Albeit a deceptively simple gas-phase system, malonaldehyde has all the salient features necessary to test the methodology development toward a description of chemical dynamics in complex systems. Extension of the approach to treat nuclear quantum effects is made possible through coupling to path integral methods as well as by the path integral quantum TST and quantum instanton formalisms. Experimental measurements of rates and KIEs are important tools of chemical kinetics for elucidating the mechanism of complex chemical reactions. In situations where the KIE is important, one can either estimate this measurable by directly computing the ratio of the isotopic rates or by TI integration with respect to mass. Our results for malonaldehyde show both approaches to give similar KIEs, although TI integration with respect to mass is computationally less expensive compared to explicit construction of the PMFs. More importantly, the above functionalities of the DG-EVB approach have been made accessible to the broader community via integration within the Amber biosimulation suite.

Some needed improvements to the DG-EVB method, nevertheless, remain. The most apparent of these is our strategy for filling in the swimming hole near the transition state region by including a UFF repulsive interaction and Amber angle and dihedral terms to the diabatic state energy. Although this practical patch appears functional, a more general solution is desirable. Using a more realistic dissociative potential beyond the harmonic approximation, such as

a Morse-type interaction, and using a full force field prescription for the H_{ii} terms will be the subjects of future development of the DG approach. It is important to emphasize that H_{11} and H_{22} need to be higher than ϵ_{ψ} for the EVB approach to function satisfactorily.²⁷ Some solutions to overcome anomalous behavior of EVB include (1) adding more ab initio data points for the fitting procedure, (2) modifying the prescription of H_{ii} , and (3) modifying the prescription of H_{12} . The DG-EVB methods development has focused on options 1 and 2, either individually or in combination, while MCOMM has explored option 3.⁴⁸ In a recent paper, Truhlar and Tishchenko allowed H_{12} to become imaginary (i.e., equivalent to our letting H_{12}^2 become negative).⁵²

Additionally, the current DG implementation requires the user to optimize the Gaussian α_K parameters such that the DG energy, gradient, and Hessian reproduce corresponding ab initio information along IRC geometries. For the case of a single DG data point, this parameter is unique. When multiple Gaussians are used to enhance the fit, the choice of α_K parameters will affect the quality of the surface. Although this parametrization overhead cannot be circumvented, additional application of the DG method may provide some *rule of thumb* intuition as to the better choices of α_K values. For example, is it necessary to assign unique values for each distributed Gaussian or will an average parameter suffice? Is it better to use α_K values that provide more diffuse Gaussians or are more localized Gaussians the optimal choice? These are questions related to parameter sensitivity analysis that will be the subject of future studies. Furthermore, an extension of DG-EVB to treat chemical reactions in condensed environments requires the development of hybrid methods that fit only part of the *active site* region to ab initio data, as electronic structure calculations of the whole condensed phase configuration are prohibitively infeasible.

Acknowledgment. The Office of Naval Research (N00014-05-1-0457) supported this research. J.L.S. is grateful for the computer time on the Wayne State University Grid. K.F.W. is grateful for an allocation of computer time from the Center for High Performance Computing at the University of Utah. A portion of the computational resources for this project have been provided by the National Institutes of Health (Grant #NCRR 1 S10 RR17214-01) on the Arches Metacluster, administered by the University of Utah Center for High Performance Computing.

Supporting Information Available: Details of the electronic structures for malonaldehyde used in the DG-EVB surface fit. This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- (1) Freddolino, P. L.; Liu, F.; Gruebele, M.; Schulten, K. *Biophys. J.* **2008**, *94*, L75.
- (2) *Petascale Computing: Algorithms and Applications*; Bader, D., Ed.; Chapman & Hall/CRC Press: Boca Raton, FL, 2008.
- (3) Maragakis, P.; Lindorff-Larsen, K.; Eastwood, M. P.; Dror, R. O.; Klepeis, J. L.; Arkin, I. T.; Jensen, M. Ø.; Xu, H.;

- Trbovic, N.; Friesner, R. A.; Palmer, A. G.; Shaw, D. E. *J. Phys. Chem. B* **2008**, *112*, 6155.
- (4) Patel, S.; Brooks, C. L. *J. Comput. Chem.* **2004**, *25*, 1.
- (5) Patel, S.; Mackerell, A. D., Jr.; Brooks, C. L. *J. Comput. Chem.* **2004**, *25*, 1504.
- (6) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. *J. Comput. Chem.* **2004**, *25*, 1157.
- (7) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. *Proteins: Struct., Funct., Bioinf.* **2006**, *65*, 712.
- (8) Xu, Z.; Luo, H. H.; Tieleman, D. P. *J. Comput. Chem.* **2007**, *28*, 689.
- (9) Soares, T. A.; Hünenberger, P. H.; Kastenholz, M. A.; Krätzler, V.; Lenz, T.; Lins, R. D.; Oostenbrink, C.; van Gunsteren, W. F. *J. Comput. Chem.* **2005**, *26*, 725.
- (10) Hill, J.-R.; Freeman, C. M.; Subramanian, L. In *Rev. Comput. Chem.*; Lipkowitz, K. B., Boyd, D. B., Eds.; Wiley-VCH: New York, 2007; p 141.
- (11) Balint-Kurti, G. G. *Adv. Chem. Phys.* **1975**, *30*, 137.
- (12) Eyring, H. *Trans. Faraday Soc.* **1938**, *34*, 3.
- (13) Evans, M. G.; Polanyi, M. *Trans. Faraday Soc.* **1938**, *34*, 11.
- (14) Ogg, R. A.; Polanyi, M. *Trans. Faraday Soc.* **1934**, *31*, 604.
- (15) Evans, M. G.; Warhurst, E. *Trans. Faraday Soc.* **1938**, *34*, 614.
- (16) Venkatnathan, A.; Voth, G. A. *J. Chem. Theory Comput.* **2005**, *1*, 36.
- (17) Warshel, A.; Weiss, R. M. *J. Am. Chem. Soc.* **1980**, *102*, 6218.
- (18) Warshel, A. *Computer Modeling of Chemical Reactions in Enzymes and Solutions*; John Wiley & Sons Inc.: New York, 1997.
- (19) Dal Peraro, M.; Ruggerone, P.; Raugei, S.; Gervasio, F. L.; Carloni, P. *Curr. Opin. Struct. Biol.* **2007**, *17*, 149.
- (20) Hutter, J.; Curioni, A. *Parallel Computing* **2005**, *31*, 1.
- (21) Carloni, P.; Rothlisberger, U.; Parrinello, M. *Acc. Chem. Res.* **2002**, *35*, 455.
- (22) Bolton, K.; Hase, W. L. In *Modern Methods for Multidimensional Dynamics*; Thompson, D. L., Ed.; World Scientific: Singapore, 1998; p 143.
- (23) Hase, W. L.; Song, K. H.; Gordon, M. S. *Comput. Sci. Eng.* **2003**, *5*, 36.
- (24) Schlegel, H. B. *Bull. Korean Chem. Soc.* **2003**, *24*, 837.
- (25) Schlegel, H. B.; Sonnenberg, J. L. *J. Chem. Theory Comput.* **2006**, *2*, 905.
- (26) Sonnenberg, J. L.; Schlegel, H. B. *Mol. Phys.* **2007**, *105*, 2719.
- (27) Sonnenberg, J. L.; Wong, K. F.; Voth, G. A.; Schlegel, H. B. *J. Chem. Theory Comput.* **2009**, *5*, 949.
- (28) Case, D. A.; Darden, T. A.; Cheatham, T. E., III; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Walker, R. C.; Zhang, W.; Merz, K. M.; Roberts, B. P.; Wang, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Kolossváry, I.; Wong, K. F.; Paesani, F.; Vaníček, J.; Wu, X.; Brozell, S. R.; Steinbrecher, T.; Gohlke, H.; Cai, Q.; Ye, X.; Wang, J.; Hsieh, M.-J.; Cui, G.; Roe, D. R.; Mathews, D. H.; Seetin, M. G.; Sagui, C.; Babin, V.; Luchko, T.; Gusarov, S.; Kovalenko, A.; Kollman, P. A. *Amber*; University of California: San Francisco, 2010.
- (29) Feynman, R. P.; Hibbs, A. R. *Quantum Mechanics and Path Integrals*; McGraw-Hill: New York, 1965.
- (30) Feynman, R. P. *Statistical Mechanics*; Benjamin: Reading, MA, 1972.
- (31) Kleinert, H. *Path Integrals in Quantum Mechanics, Statistics, and Polymer Physics*; World Scientific: Singapore, 1995.
- (32) Miller, W. H. *J. Chem. Phys.* **1975**, *62*, 1899.
- (33) Miller, W. H.; Zhao, Y.; Ceotto, M.; Yang, S. *J. Chem. Phys.* **2003**, *119*, 1329.
- (34) Yamamoto, T.; Miller, W. H. *J. Chem. Phys.* **2004**, *120*, 3086.
- (35) Miller, W. H.; Schwartz, S. D.; Tromp, J. W. *J. Chem. Phys.* **1983**, *79*, 4889.
- (36) Wang, Y.; Braams, B. J.; Bowman, J. M.; Carter, S.; Tew, D. P. *J. Chem. Phys.* **2008**, *128*, 224314.
- (37) Viel, A.; Coutinho-Neto, M. D.; Manthe, U. *J. Chem. Phys.* **2007**, *126*, 024308.
- (38) Mil'nikov, G. V.; Yagi, K.; Taketsugu, T.; Nakamura, H.; Hirao, K. *J. Chem. Phys.* **2004**, *120*, 5036.
- (39) Mil'nikov, G. V.; Yagi, K.; Taketsugu, T.; Nakamura, H.; Hirao, K. *J. Chem. Phys.* **2003**, *119*, 10.
- (40) Yagi, K.; Taketsugu, T.; Hirao, K. *J. Chem. Phys.* **2001**, *115*, 10647.
- (41) Hazra, A.; Skone, J. H.; Hammes-Schiffer, S. *J. Chem. Phys.* **2009**, *130*, 054108.
- (42) Ben-Nun, M.; Martinez, T. *J. Phys. Chem. A* **1999**, *103*, 6055.
- (43) Sewell, T. D.; Guo, Y.; Thompson, D. L. *J. Chem. Phys.* **1995**, *103*, 8557.
- (44) Shida, N.; Barbara, P. F.; Almolöf, J. E. *J. Chem. Phys.* **1989**, *91*, 4061.
- (45) Makri, N.; Miller, W. H. *J. Chem. Phys.* **1989**, *91*, 4026.
- (46) Ruf, B. A.; Miller, W. H. *J. Chem. Soc., Faraday Trans.* **1988**, *84*, 1523.
- (47) Carrington, T.; Miller, W. H. *J. Chem. Phys.* **1986**, *84*, 4364.
- (48) Higashi, M.; Truhlar, D. G. *J. Chem. Theory Comput.* **2008**, *4*, 790.
- (49) Albu, T. V.; Corchado, J. C.; Truhlar, D. G. *J. Phys. Chem. A* **2001**, *105*, 8465.
- (50) Kim, Y.; Cochado, J. C.; Villa, J.; Xing, J.; Truhlar, D. G. *J. Chem. Phys.* **2000**, *112*, 2718.
- (51) Lin, H.; Pu, J.; Albu, T. V.; Truhlar, D. G. *J. Phys. Chem. A* **2004**, *108*, 4112.
- (52) Tishchenko, O.; Truhlar, D. G. *J. Chem. Theory Comput.* **2009**, *5*, 1454.
- (53) Chandler, D. *Introduction to Modern Statistical Mechanics*; Elsevier: New York, 1991.
- (54) Chang, Y.-T.; Miller, W. H. *J. Phys. Chem.* **1990**, *94*, 5884.
- (55) Chang, Y.-T.; Minichino, C.; Miller, W. H. *J. Chem. Phys.* **1992**, *96*, 4341.
- (56) Minichino, C.; Voth, G. A. *J. Phys. Chem. B* **1997**, *101*, 4544.
- (57) Saad, Y.; Schultz, M. H. *SIAM J. Sci. Stat. Comput.* **1986**, *7*, 856.
- (58) Pulay, P. *Chem. Phys. Lett.* **1980**, *73*, 393.
- (59) Pulay, P. *J. Comput. Chem.* **1982**, *3*, 556.

- (60) Pulay, P. In *Molecular Quantum Mechanics: Analytic Gradients and Beyond*; Csaszar, A. G., Fogarasi, G., Schaefer, H. F., III, Szalay, P. G., Eds.; ELTE Institute of Chemistry: Budapest, 2007; p 71.
- (61) Arndt, M.; Nairz, O.; Vos-Andreae, J.; Keller, C.; van der Zouw, G.; Zeilinger, A. *Nature* **1999**, *401*, 680.
- (62) Zuev, P. S.; Sheridan, R. S.; Albu, T. V.; Truhlar, D. G.; Hrovat, D. A.; Borden, W. T. *Science* **2003**, *299*, 867.
- (63) McMahon, R. J. *Science* **2003**, *299*, 833.
- (64) Chandler, D.; Wolynes, P. G. *J. Chem. Phys.* **1981**, *74*, 4078.
- (65) Ceperley, D. M. *Rev. Mod. Phys.* **1995**, *67*, 279.
- (66) Berne, B. J.; Thirumalai, D. *Annu. Rev. Phys. Chem.* **1986**, *37*, 401.
- (67) Martyna, G. J.; Klein, M. L.; Tuckerman, M. *J. Chem. Phys.* **1992**, *97*, 2635.
- (68) Hänggi, P.; Talkner, P.; Borkovec, M. *Rev. Mod. Phys.* **1990**, *62*, 251.
- (69) Kapral, R.; Consta, S.; McWhirter, L. In *Classical and Quantum Dynamics in Condensed Phase Simulations*; Berne, B. J., Ciccotti, G., Coker, D. F., Eds.; World Scientific: Singapore, 1998.
- (70) Hinsen, K.; Roux, B. *J. Chem. Phys.* **1997**, *106*, 3567.
- (71) Roux, B. *Comput. Phys. Commun.* **1995**, *91*, 275.
- (72) Voth, G. A.; Chandler, D.; Miller, W. H. *J. Chem. Phys.* **1989**, *91*, 7749.
- (73) Yamamoto, T.; Miller, W. H. *J. Chem. Phys.* **2005**, *122*, 044106.
- (74) Vaníček, J.; Miller, W. H. *J. Chem. Phys.* **2007**, *127*, 114309.
- (75) Cao, J.; Berne, B. J. *J. Chem. Phys.* **1989**, *91*, 6359.
- (76) Kumar, S.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A.; Rosenberg, J. M. *J. Comput. Chem.* **1992**, *13*, 1011.
- (77) Kumar, S.; Rosenberg, J. M.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A. *J. Comput. Chem.* **1995**, *16*, 1339.
- (78) Parthiban, S.; Martin, J. M. L. *J. Chem. Phys.* **2001**, *114*, 6014.
- (79) Barnes, E. C.; Petersson, G. A.; Montgomery, J. A., Jr.; Frisch, M. J.; Martin, J. M. L. *J. Chem. Theory Comput.* **2009**, *5*, 2687.
- (80) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Scalmani, G.; Mennucci, B.; Barone, V.; Petersson, G. A.; Caricato, M.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Li, X.; Hratchian, H. P.; Peralta, J. E.; Izmaylov, A. F.; Kudin, K. N.; Heyd, J. J.; Brothers, E.; Staroverov, V.; Zheng, G.; Kobayashi, R.; Normand, J.; Sonnenberg, J. L.; Ogliaro, F.; Bearpark, M.; Parandekar, P. V.; Ferguson, G. A.; Mayhall, N. J.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Burant, J. C.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; 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.; Chen, W.; Wong, M. W.; Pople, J. A. *Gaussian 03*, Revision G.01; Gaussian, Inc, Wallingford, CT, 2007.
- (81) Rappé, A. K.; Casewit, C. J.; Colwell, K. S.; Goddard, W. A.; Skiff, W. M. *J. Am. Chem. Soc.* **1992**, *114*, 10024.
- (82) Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids*; Clarendon Press: Oxford, U. K., 1987.
- (83) Pastor, R. W.; Brooks, B. R.; Szabo, A. *Mol. Phys.* **1988**, *65*, 1409.
- (84) Tuckerman, M.; Marx, D. *Phys. Rev. Lett.* **2001**, *86*, 4946.
- (85) Iftimie, R.; Schofield, J. *J. Chem. Phys.* **2001**, *115*, 5891.
- (86) Iftimie, R.; Schofield, J. *J. Chem. Phys.* **2001**, *114*, 6763.

CT900579K

JCTC

Journal of Chemical Theory and Computation

Rigid-Body Molecular Dynamics of Fullerene-Based Nanocars on Metallic Surfaces

Sergei S. Konyukhov,[†] Ilya V. Kupchenko,[†] Alexander A. Moskovsky,[†]
Alexander V. Nemukhin,^{†,‡} Alexey V. Akimov,[§] and Anatoly B. Kolomeisky^{*,§}

Department of Chemistry, M.V. Lomonosov Moscow State University, Leninskie Gory 1/3, Moscow 119991, and N.M. Emanuel Institute of Biochemical Physics, Russian Academy of Sciences, ul. Kosygina 4, Moscow 119994, Russian Federation, and Department of Chemistry, Rice University, Houston, Texas 77005

Received February 19, 2010

Abstract: Methodical problems of coarse-grained-type molecular dynamics, namely, rigid-body molecular dynamics (RB MD), are studied by investigating the dynamics of nanosized molecular vehicles called nanocars that move on gold and silver surfaces. Specifically, we analyzed the role of thermostats and the effects of temperature, couplings, and correlations between rigid fragments of the nanocar molecule in extensive RB MD simulations. It is found that the use of the Nosé–Poincaré thermostat does not introduce systematic errors, but the time trajectories might be required to be limited to not accumulate large numerical integration errors. Correlations in the motion of different fragments of the molecules are also analyzed. Our theoretical computations also point to the importance of temperature, interfragment interactions, and interactions with surfaces and to the nature of the surface for understanding mechanisms of motion of single-molecule transporters.

1. Introduction

Computer simulations based on molecular dynamics (MD) methods are critically important for understanding mechanisms of fundamental processes. MD is a powerful and convenient tool for analyzing different biological and chemical systems such as protein folding phenomena, motor protein dynamics, and transport across channels and membranes. Recently, MD methods have also been applied for studying artificial molecular motors, rotors,^{1–8} and single-molecule transporters that include nanocars.^{9–17}

In principle, classical MD methods are based on solving equations of motion for every atom of the system at all times, and this allows visualization and prediction of the dynamics of the system at the molecular level. However, a real system has a very large number of particles and degrees of freedom, and it becomes prohibitively expensive to solve these

equations and to get a reasonable description of the system. It is known that full atomic MD simulations cannot capture processes that involve large groups of atoms for times longer than hundreds of nanoseconds, while typical chemical and biological time scales range from 1 ms to 1 min. In addition, in many systems some degrees of freedom are less important than others for dynamical properties. These considerations stimulated a development of the so-called coarse-grained MD methods in which groups of atoms are viewed as new “effective” particles and some degrees of freedom are ignored. These approaches significantly accelerate computations by decreasing the number of particles and/or degrees of freedom. However, this raises an important question on how realistic these coarse-grained descriptions of the studied phenomenon are.

The goal of this study is to analyze the applicability of a version of coarse-grained MD for analyzing single-molecule dynamics. Specifically, we will study recently developed rigid-body molecular dynamics (RB MD) methods^{18–24} that have been successfully applied for studying complex dynamics of artificial molecular motors (nanocars)¹⁷ and molecular rotors.^{25–29} Using this approach, many important details of

* Corresponding author phone: (713) 348-5672; e-mail: tolya@rice.edu.

[†] M.V. Lomonosov Moscow State University.

[‡] Russian Academy of Sciences.

[§] Rice University.

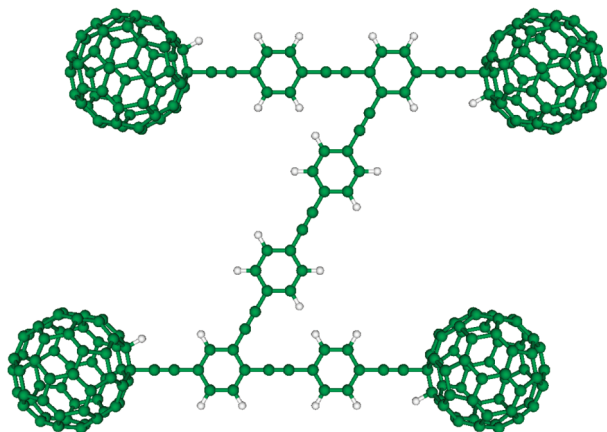


Figure 1. Model of the z-car with four fullerene wheels chemically bound to a chassis.

the molecular motion at the nanoscale have been uncovered. It also allowed us to better understand recent single-molecule measurements on molecular motors and rotors. In RB MD the molecules are viewed as collections of rigid fragments; i.e., many vibrational and rotational degrees of freedom are neglected. Since RB MD provides a reasonable description of the dynamics of nanocars on metallic surfaces, as compared with experiments, we have chosen this system as our testing ground for checking the reliability of RB MD calculations. By performing extensive numerical simulations, we analyze the effect of using a specific Nosé–Poincaré thermostat^{30–34} and the effect of temperature, correlations, molecular flexibility, and interactions between different parts of the system. There are several types of nanocar molecules that have been synthesized in recent years. In our work the dynamics of a single molecule, known as a z-car and shown in Figure 1, is studied on gold and silver surfaces because of the high stability of this molecule on these surfaces.

It should be noted that the development of a realistic RB MD tool is also very important for practical applications. It will allow us to better understand mechanisms of motion on the surface by single-molecule transporters, and it will help to create efficient nanoscale transport systems capable of performing useful work at the molecular level.

2. Methods

To test the RB MD method, we used a simplified model of the dynamics of the z-car molecule (illustrated in Figure 1) on gold and silver surfaces under conditions close to vacuum which are similar to what is observed in experiments.¹² The original z-car molecule has several tail fragments that increase its solubility in organic solvents and make the molecule more rigid. In our computations we used a smaller molecule without tails to have a more comprehensive description of the system dynamics. In further discussions, we call this simplified version a z-car molecule.

The molecules move along the metal surface, which is viewed as an ideal two-dimensional square-lattice plane with the lattice parameters equal to 4.07 Å for gold and 4.08 Å for silver. Our main assumption here is that the metal lattice is rigid and interactions with the z-car do not change the underlying lattice geometry of the surface. Effectively, this

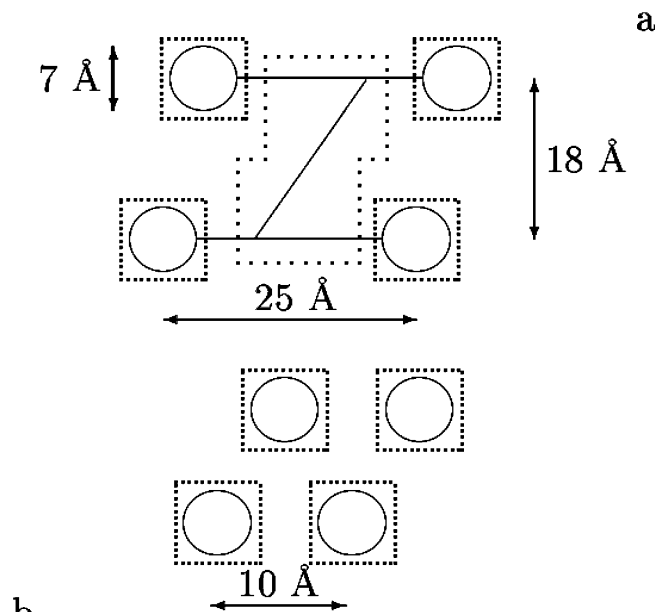


Figure 2. Partitioning the model systems into fragments: (a) z-car, (b) four wheels ($4 \times C_{60}$).

corresponds to taking into account only the upper surface atom layer although the effect of the bottom levels has been analyzed as well.

The structure of the z-car molecule suggests that it is reasonable to divide it into five rigid segments: four corresponding to the wheels and one to the chassis. Since we are interested in correlations and interactions between different parts of the z-car molecule, for comparison we also analyzed the molecular dynamics of a separate system consisting of only four fullerene wheels moving on the same surfaces. Schematic pictures of the considered particles and their geometries are presented in Figure 2.

The crucial part of any MD calculation is the potential energy. Most of the parameters for the present simulations have been taken from the CHARMM27 force field^{35,36} augmented by those from the UFF parameter set.³⁷ Interactions with silver atoms have been parametrized³⁸ to reproduce the adsorption energies of some relatively small molecules, such as fullerenes. Since charge transfer processes might play a significant role in interactions between fullerenes and metallic surfaces,^{39–43} this procedure allowed us to take into account, at least partially, various chemisorption effects. The overall potential energy is additive, and it can be written in the following form:

$$E = E_{\text{bond}} + E_{\text{angle}} + E_{\text{vdw}} + E_{\text{Coulomb}} + E_{\text{surface-molecule}} \quad (1)$$

The first term describes the radial part of the covalent bond interaction:

$$E_{\text{bond}} = \sum_{\text{bonds}} kb_i(r_i - r_i^{\text{eq}})^2 \quad (2)$$

where r_i and r_i^{eq} are the length and the equilibrium length, respectively, of bond i . The second term in eq 1 corresponds to the angular contribution to the interactions:

$$E_{\text{angle}} = \sum_{\text{angles}} ka_i(\varphi_i - \varphi_i^{\text{eq}})^2 \quad (3)$$

where φ_i and φ_i^{eq} are the angle and the equilibrium angle, respectively, between two neighboring bonds. In our calculations the effect of torsion interactions has been mostly ignored. The third term in eq 1 describes the noncovalent interactions in the system that are assumed to be of the standard Lennard-Jones type:

$$E_{\text{vdw}} = \sum_{ij} \varepsilon_{ij} \left[\left(\frac{r_{ij}^0}{r_{ij}} \right)^{12} - 2 \left(\frac{r_{ij}^0}{r_{ij}} \right)^6 \right] \quad (4)$$

where $\varepsilon_{ij} = (\varepsilon_i \varepsilon_j)^{1/2}$ and r_{ij} is the distance between atoms i and j . The fourth term in eq 1 corresponds to electrostatic interactions:

$$E_{\text{Coulomb}} = \sum_{i < j} \left(\frac{q_i q_j}{r_{ij}} \right) \quad (5)$$

where q_i is the charge on atom i . This contribution was neglected for the case of the Au surface. However, the effect of the charge transfer in the case of the Ag surface was implicitly taken into account in the utilized potential.³⁸ Finally, the last term represents the interaction between the molecule and the surface. For the Au surface it was taken to be of the Lennard-Jones type, while for the Ag surface a Morse potential has been used. In our calculations on the gold surface, we explicitly considered noncovalent interactions with the 625 closest surface gold atoms. Note also that we took into account van der Waals interactions between different parts of the z-car molecule. For the silver surface we considered explicitly interactions with the 2134 closest surface atoms.

MD simulations have been performed with the help of the Nosé–Poincaré thermostat^{30–34} at $T = 300$ K for dynamics on the gold surface and for the range of temperatures between 200 and 600 K for the motion of z-car molecules on the silver surface. We have used for simulation two computer codes that were independently developed in our groups. The length of each trajectory was 20 ns for the dynamics on gold and 1 ns for the silver surface with a time step of 1 fs. All obtained dynamic properties have been averaged over 48 trajectories for Au and over 10–30 trajectories (at each temperature) for the Ag surface.

3. Results and Discussion

3.1. Evaluation of the Thermostat. A crucial element in our RB MD simulations is utilization of the Nosé–Poincaré thermostat, and it is important to understand its advantages and limitations. One of the special properties of this thermostat that distinguishes it from other thermostats is that the dynamics produced with the help of the Nosé–Poincaré thermostat is Hamiltonian. This observation allowed us to control, at least partially, the quality of numerical integrations of equations of motion by monitoring a special function called the Nosé–Poincaré invariant, H_{N-P} , which corresponds exactly to a Hamiltonian function of another system that approximates the dynamics in the given system. The quality

of the calculations can be judged by analyzing the invariant as a function of time. Deviations can be approximated as a linear function:

$$H_{N-P}(t) = At + B \quad (6)$$

The dependence of invariants as a function of time for MD simulations on a gold surface is shown in Figure 3. In addition, Table 1 provides information, averaged over all trajectories, for the absolute values of H_{N-P} and also for coefficients A and B . It can be seen that all these parameters are similar in value with absolute errors, and one might conclude that the algorithm of numerical integration of the equations of motion used in our RB MD simulations does not produce systematic errors. In addition, it could be noted that if the trend in the invariant does not change, then we can estimate the maximal allowed time length, t_{max} , of trajectories in our method by comparing the deviations of the invariant and typical interaction energies in the system (~ 1 kcal/mol). In this system it is equal to ~ 10 ns. Thus, there are situations when longer trajectories might not be as precise in measuring system dynamics. However, this observation strongly depends on the time-dependent behavior of the invariant, and it should be carefully checked in all systems. Our calculations on a Ag surface suggest that preliminary equilibration of the structure might decrease the trend in the invariant, and longer trajectories (up to 1 ms) could be used for analyzing the dynamics of nanocars. It should be noted also that in our analysis of correlations we included the initial time segments to have the most conservative estimate of the parameters of the system. Removing the initial times might significantly improve the computer simulation results.

Another way of checking the quality of the utilized thermostat is to analyze how it reproduces the properties of the canonical ensemble. This is related to the problem of ergodicity for trajectories obtained with the help of the Nosé–Poincaré thermostat. It is convenient to view the kinetic energy of the system as a proper characteristic property for reproducing canonic ensemble properties. The kinetic energy can be written as the sum of translational and rotational contributions for each fragment:

$$E_{\text{kin}} = \sum_{\text{fragments}} (E_{\text{kin}}^{\text{trans}} + E_{\text{kin}}^{\text{rot}}) \quad (7)$$

In this equation we have

$$E_{\text{kin}}^{\text{trans}} = \frac{1}{2m} (p_{\text{trans},x}^2 + p_{\text{trans},y}^2 + p_{\text{trans},z}^2) \quad (8)$$

where m is the mass of the fragment and $p_{\text{trans},i}$ are projections of the translational momentum on axis i . Similarly for the rotational kinetic energy

$$E_{\text{kin}}^{\text{rot}} = \frac{p_{\text{rot},x}^2}{2I_x} + \frac{p_{\text{rot},y}^2}{2I_y} + \frac{p_{\text{rot},z}^2}{2I_z} \quad (9)$$

where $p_{\text{rot},i}$ are projections of rotational momentum and $I_{x,y,z}$ are the moments of inertia in the system. In the canonical

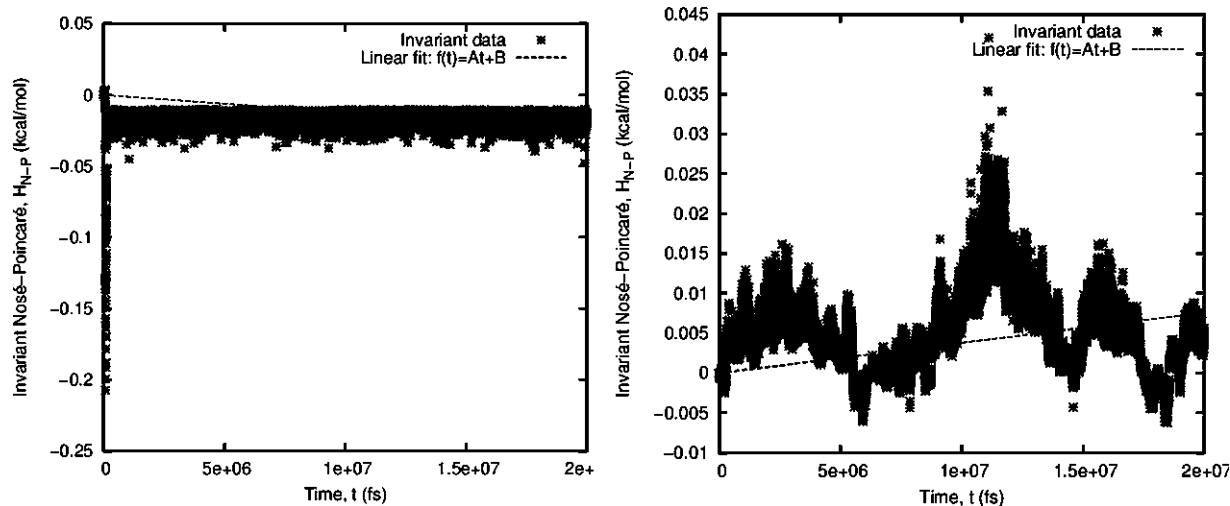


Figure 3. Time dependence of the Nosé–Poincaré invariant for $4 \times C_{60}$ (left) and z-car (right) on a gold surface.

Table 1. Nosé–Poincaré Invariant Values, H_{N-P} , and the Coefficients of the Linear Approximation (Eq 6) of the Trend

system	$H_{N-P} \pm \Delta H_{N-P}$, kcal/mol	$A \pm \Delta A$, (kcal/mol) fs^{-1}	$B \pm \Delta B$, kcal/mol
$4 \times C_{60}$	$(4 \pm 2) \times 10^{-2}$	$(-3 \pm 5) \times 10^{-8}$	$(4 \pm 2) \times 10^{-2}$
z-car	$(3 \pm 2) \times 10^{-3}$	$(3 \pm 2) \times 10^{-7}$	$(-1 \pm 1) \times 10^{-4}$

ensemble the probabilities for all moments follow a Boltzmann distribution:

$$f(p) = A_f p_r^2 \exp\left[-\frac{p_r^2}{B_f}\right] \quad (10)$$

where p_r is the reduced moment given by

$$p_r = \frac{p_{trans,j}}{m} \quad \text{or} \quad p_r = \frac{p_{rot,j}}{\sqrt{I_j}} \quad j = x, y, z \quad (11)$$

and the coefficients A_f and B_f are equal to

$$A_f = \frac{4}{\sqrt{\pi}} \left(\frac{1}{2k_B T}\right)^{-3/2} \quad B_f = 2k_B T \quad (12)$$

The distributions obtained in our MD simulations are compared with distributions theoretically calculated via eq 10 in Figures 4 and 5.

The results of our simulations reproduce quite well theoretical distributions for the system of four fullerene wheels and for the wheels in the z-car. However, some deviations are observed for the chassis of the z-car. Most probably this is related to deviations from ergodicity for the results obtained with the Nosé–Poincaré thermostat. The chassis of the z-car molecules is strongly limited in its motion by coupling to the fullerene wheels, and probably it does not help to explore fully the phase space of the system to satisfy the Boltzmann distribution.

3.2. Correlations in the Motion of Fullerene Wheels.

Our previous studies have shown that rotation of the wheels is a dominating factor in the mobility of nanocars.²⁴ To understand the mechanism of the motion of these molecules, it is important to estimate correlations in the wheel rotations.

To do this, we compared the motion of the z-car molecule with the dynamics of a system consisting of only four fullerene wheels. Our MD simulations indicate that if we position four fullerene wheels like in the z-car molecule, then after some time they come together and stay as one dynamic cluster with a distance between neighboring fullerenes on the order of 1 nm. This is due to van der Waals interactions between the particles. For the z-car, these interactions between the wheels are very weak and fullerenes do not directly affect each other. Any possible correlations are the result of the effective interactions with the whole system.

To quantify rotational correlations, we introduce several functions that we might call correlation functions. Specific values for these functions, as described below, will provide a relative measure of the correlations. First, we define a function C

$$C = \frac{2}{3} \frac{\sum_{i < j} (\mathbf{w}_i \mathbf{w}_j)}{\sum_i (\mathbf{w}_i \mathbf{w}_i)} \quad (13)$$

where $(\mathbf{w}_i \mathbf{w}_j)$ is a scalar product of two angular velocity vectors for wheels i and j . The physical meaning of this function is the following. If the absolute values and directions of all angular velocities are the same, i.e., the case of very strong correlations, the value of this function is equal to 1. If the motion of the wheel is uncorrelated, the value of the function C should be between 0 and $-1/3$. The last value, for example, can be obtained in the case where the absolute values of all angular velocities are the same and two vectors are in the upper direction while two other vectors are in the opposite direction.

The function C provides a measure of both correlations in the direction and in the absolute values of the rotational speeds of the wheels. We also introduce two other functions, C_{dir} and C_{abs} , that are more specific indicators of the corresponding correlations:

$$C_{dir} = \frac{1}{4} \sum_{i=1}^4 (\mathbf{n}_i \mathbf{n}_{av}) \quad (14)$$

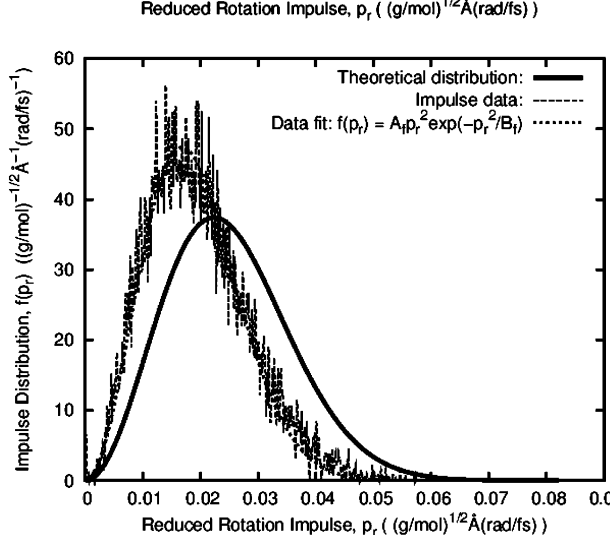
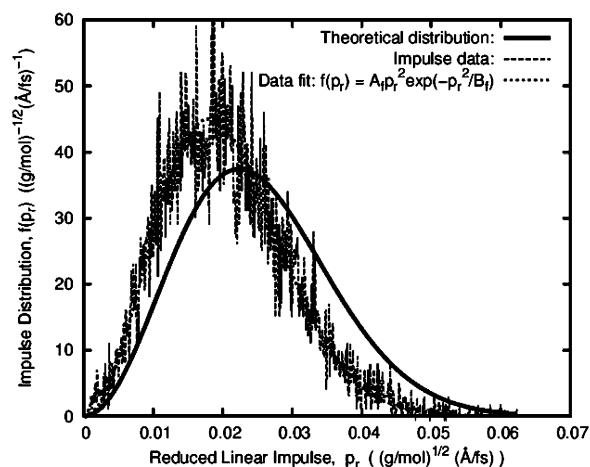
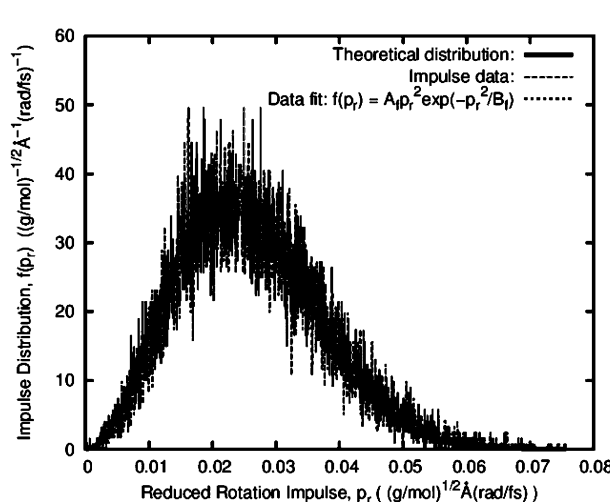
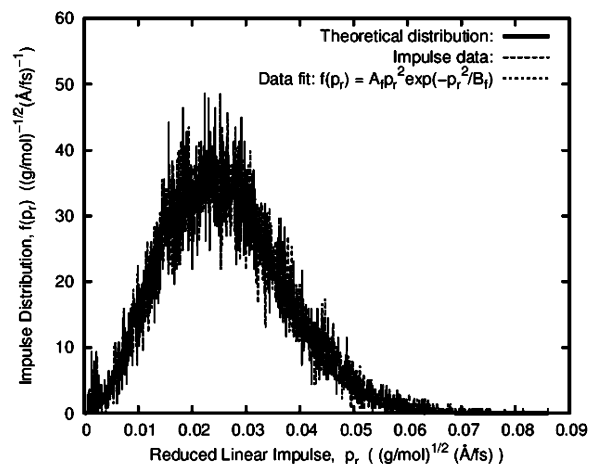
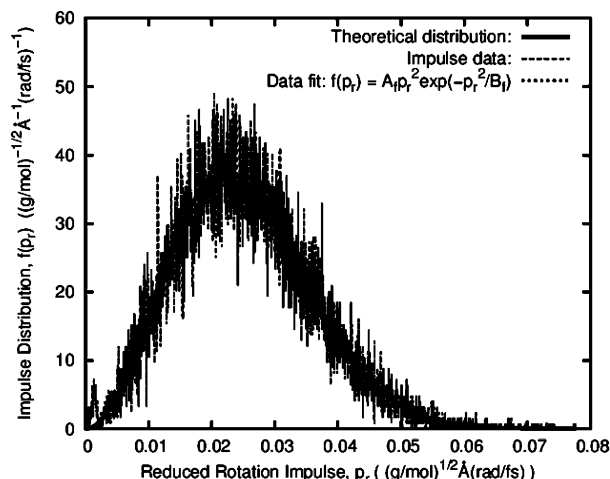
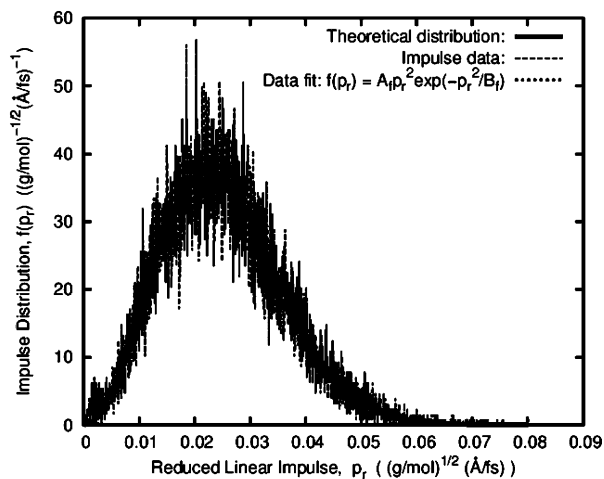


Figure 4. Distributions of the reduced linear impulses: 4 × C₆₀ (top), z-car wheels (middle), z-car chassis (bottom).

with

$$\mathbf{n}_i = \frac{\mathbf{w}_i}{|\mathbf{w}_i|} \quad \mathbf{n}_{av} = \frac{\sum_{i=1}^4 \mathbf{n}_i}{|\sum_{i=1}^4 \mathbf{n}_i|} \quad (15)$$

The possible numerical values for this function range from 0 (no correlations) to 1 (very strong correlations). The correlations in the absolute values of the angular velocities can be measured with the help of another function:

Figure 5. Distributions of the reduced rotational impulses: 4 × C₆₀ (top), z-car wheels (middle), z-car chassis (bottom).

$$C_{abs} = \frac{\sqrt{\sum_{i=1}^4 (|\mathbf{w}_i| - \overline{|\mathbf{w}|})^2}}{|\overline{|\mathbf{w}|}|} \quad \overline{|\mathbf{w}|} = \frac{1}{4} \sum_{i=1}^4 |\mathbf{w}_i| \quad (16)$$

Typical temporal dependencies of the correlation functions are presented in Figures 6–8, and the averaged distributions of the values are given in Tables 2 and 3. All functions start from 0 since at $t = 0$ it is assumed that there are no

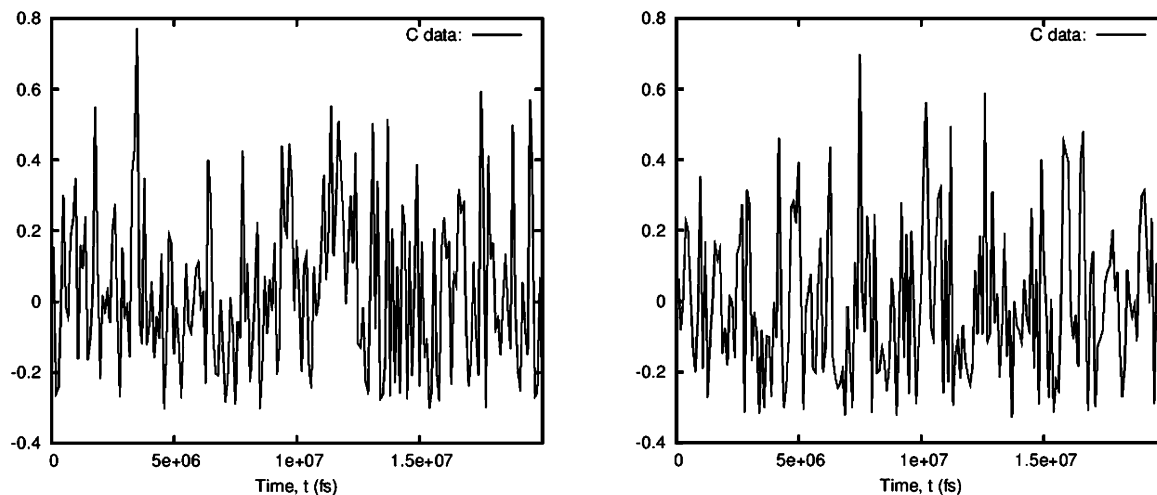


Figure 6. Typical temporal dependences of the correlation function C (eq 13) for the $4 \times C_{60}$ system (left) and for the z-car (right).

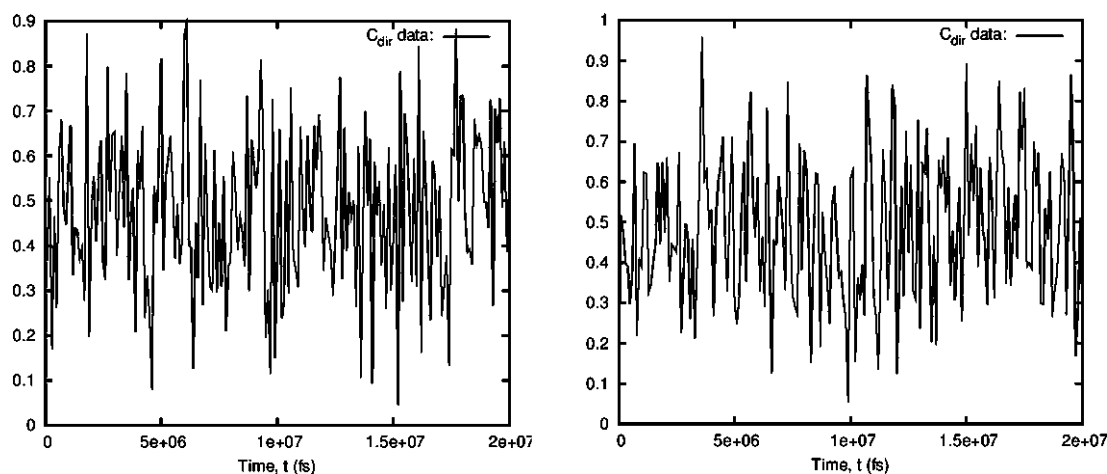


Figure 7. Typical temporal dependences of the correlation function C_{dir} (eq 14) for the $4 \times C_{60}$ system (left) and for the z-car (right).

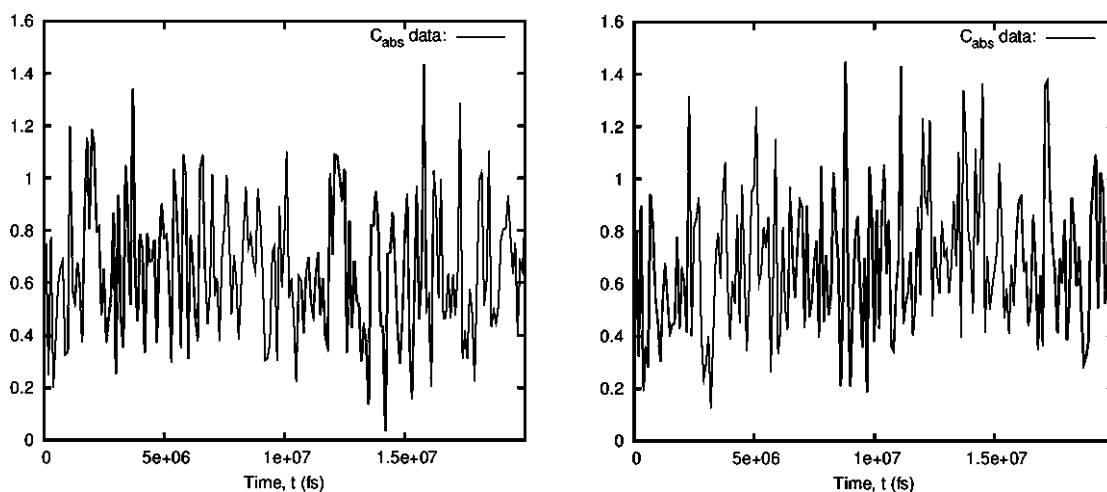


Figure 8. Typical temporal dependences of the correlation function C_{abs} (eq 16) for the $4 \times C_{60}$ system (left) and for the z-car (right).

correlations. It is interesting to note that correlations in the wheels of the z-car molecule are similar to those of four separate fullerene wheels. Our results indicate that correlations in the motion of the wheels of the nanocars are small but not negligible. This is the result of complex interactions

between different fragments of the molecule and interactions with the surface.

3.3. Mobility of Nanocars. Experimental and theoretical studies indicate that nanocars move stochastically along the surfaces. Their mobility could be quantized by analyzing

Table 2. Distribution C (Eq 13) for the Systems $4 \times C_{60}$ and z-car

system	$0 \leq C < 1/4$	$1/4 \leq C < 1/2$	$1/2 \leq C < 3/4$	$3/4 \leq C < 1$
$4 \times C_{60}$	0.121	0.452	0.356	0.071
z-car	0.121	0.468	0.344	0.067

Table 3. Distributions C_{dir} (Eq 14) and C_{abs} (Eq 16) for the Systems $4 \times C_{60}$ and z-car

system	C_{dir}	$\Delta C_{dir}/C_{dir}, \%$	C_{abs}	$\Delta C_{abs}/C_{abs}, \%$
$4 \times C_{60}$	0.469	0.4	0.679	0.4
z-car	0.465	0.4	0.686	0.4

Table 4. Comparison of Translational and Rotational Diffusion Coefficients

system	$D_{trans}, \text{\AA}^2 \text{ fs}^{-1}$	$\Delta D_{trans}/D_{trans}, \%$	$D_{rot}, \text{rad}^2 \text{ fs}^{-1}$	$\Delta D_{rot}/D_{rot}, \%$
$4 \times C_{60}$	0.010	7	1.3×10^{-5}	12
z-car	0.015	10	2.2×10^{-5}	14

effective diffusion coefficients. In our RB MD simulations we measured rotational diffusion following the change of the angle of rotation around the axis perpendicular to the surface:

$$D_{rot} = \frac{\varphi_{max}^2}{4t_{max}} \quad (17)$$

where $t_{max} = 2 \times 10^7$ fs. In a similar way, we can estimate the translational diffusion constants for different particles:

$$D_{trans} = \frac{r_{max}^2}{4t_{max}} \quad (18)$$

where $t_{max} = 2 \times 10^7$ fs and r_{max} is the position of the particle at the end of the trajectory, assuming that it started at the origin. Rotational and translational diffusion constants for different surfaces are presented in Table 4. Comparing the results for the z-car molecule and for the cluster of four fullerene molecules, which are quite close, one could argue that the most important factor controlling the dynamics of nanocars is their interactions with the surface and overcoming the barriers created by this potential energy and not interfragment interactions. However, we should be careful in this situation since in this system interactions between different parts of the molecule are relatively small. However, this conclusion might change for stronger interfragment interactions.

3.4. Effect of Molecular Flexibility. Since in the RB MD method it is possible to divide the molecule into different sets of fragments, we can investigate the question of how molecular flexibility affects the overall dynamics. For the motion of the z-car on the silver surface we analyzed five different ways of fragmentation for the same molecule. Our most flexible molecule has five fragments (four wheels and a chassis), while in other cases we combined some wheels into one rigid fragment with the chassis. Thus, we analyzed molecules with four, three, two, one, and zero rotating wheels. Translational diffusion coefficients, computed as discussed above, as functions of the number of fragments

are shown in Figure 9. It can be seen that the diffusion constant is mostly an increasing function of the number of wheels that can rotate, although the results are quite noisy. This suggests that the rotation of fullerene wheels makes a significant contribution to the overall mobility of the z-car molecule. This can be understood using the following arguments. If all wheels are frozen, when the molecule is viewed as one rigid fragment, the only way it can move is by hopping on the surface. Allowing some wheels to rotate increases the overall mobility.

We also investigated the effect of the temperature on the mobility of the z-car molecule on the silver surface by performing RB MD at different temperatures. The results have been fitted in the standard Arrhenius expression

$$D = D_0 \exp\left[-\frac{E_a}{RT}\right] \quad (19)$$

where E_a is an effective activation energy. The activation energy as a function of the number of fragments has a nonmonotonic dependence: first it increases strongly, and then it slowly decreases (Figure 9a). We can propose the following explanation for this complex behavior. The smallest activation barrier is observed for the molecule without rotating wheels. When some wheels start to rotate, the molecule can adjust itself on the surface and find a lower minimum on the potential surface, something that the totally rigid molecule cannot do. Then it is harder for this molecule with few rotating wheels to hop because it takes more energy to overcome the larger barrier. However, increasing the number of rotating wheels also increases the contribution of rotations in the translational motion, which have smaller activation barriers.

Figure 9b also shows the effect of increasing the number of fragments of the nanocar molecule on the pre-exponential term in eq 19. Increasing the molecular flexibility strongly increases this parameter. These results can be easily understood. The pre-exponential factor describes the attempt frequency, and the more flexible molecule explores a larger phase space, allowing more attempts for the molecule to diffuse along the surface.

3.5. Comparison of Different Surfaces. In our simulations we have investigated the motion of nanocars on gold and silver surfaces. The dynamics in both cases are similar, but there are many differences. The potential used in the case of the gold surface has been obtained from CHARMM and from UFF force fields, while in the case of silver a combination of UFF and an empirically determined potential³⁸ has been used. The obtained parameters in both cases are comparable, with the exception of one. The equilibrium angle in the connection between the fullerene wheel and the chassis for the Ag surface was taken to be 120° . This is because the aromatic carbon has been utilized in these simulations on the silver surface. However, on the Au surface this angle was taken as 180° , corresponding to sp hybridization of the carbon atom. This difference leads to slightly different structures: the z-car molecule on the silver surface has a bend in the joint region of the wheels and chassis, while on the gold surface there is no such bend. As a result, the wheels are interacting stronger with the silver surface,

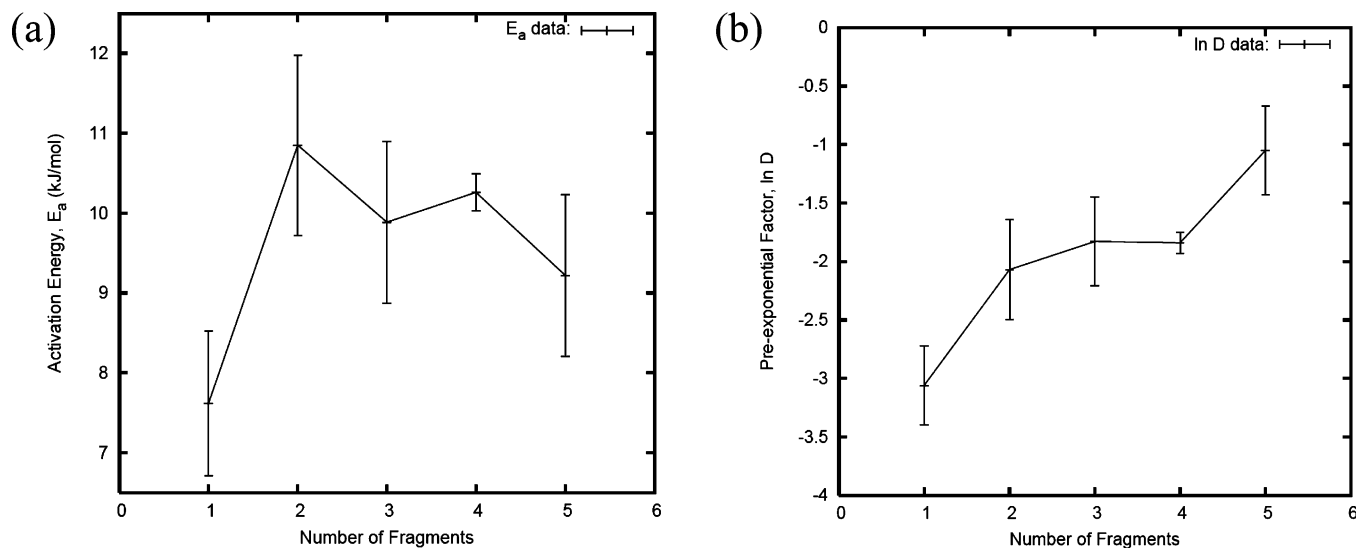


Figure 9. Activation energies (left) and pre-exponential factors (right) for diffusion of the z-car as a function of its flexibility.

leading to smaller diffusion coefficients. It will be interesting to test experimentally whether these theoretically predicted differences in the structures of the nanocars are real.

It is important to discuss how the electrostatic charges on a metal surface might influence our MD simulations. As was discussed above, to describe Ag–fullerene interactions, we have used an empirically found potential.³⁸ Since this potential provides reasonable adsorption energies for a series of molecules on the silver surface, the effect of charges is already implicitly included in our calculations. For fullerene–Au interactions the charge transfer plays an important role, and in general, it cannot be neglected. However, to not make our calculations very complex, we have neglected the electrostatic effects in this case because we are interested in more qualitative features of the dynamics that can be obtained without this contribution.

3.6. Comparison to the All-Atom MD Models. It is instructive to estimate the effects of the rigid-body restrictions on the dynamical behavior in these systems by comparing our results with those of the standard all-atom MD simulations. First, we reproduced the data from the work of Yoon et al.⁴⁴ in which the diffusion of gold clusters on graphite surfaces has been modeled. We considered one of the model systems, namely, the 140-atom gold cluster on a graphite surface (111), and precisely the same interaction potential parameters as in the original paper.⁴⁴ The diffusion coefficients have been computed in the all-atom MD simulations in complete agreement with the values reported in ref 44 and then for the modified model system in which the metal cluster was treated as a rigid body. We found that in the latter case the computed diffusion coefficients have been calculated to be approximately 1 order of magnitude larger than in the all-atom case. Qualitatively the same conclusion on larger values of diffusion coefficients estimated within the rigid-body MD simulations compared to the all-atom MD simulations has been obtained for another model system, the single C_{60} molecule on the Au(111) surface. Calculations of trajectories and estimates of the translational diffusion coefficients for the RB dynamics have been carried out as for other fullerene-based systems described above. To

perform all-atom MD simulations, we considered five layers of gold atoms with the periodical cell of $20 \times 20 \times 100 \text{ \AA}$. The lowest layer has been frozen, and the four upper levels were allowed to be flexible as consistent with the Lennard-Jones parameters for Au–Au interactions $\epsilon = 10.6 \text{ kcal/mol}$ and $\sigma = 2.57 \text{ \AA}$. The MM3 force field parameters were used to describe the all-atom C_{60} molecule. The obtained diffusion coefficient was approximately 2 orders of magnitude smaller than that estimated in the rigid-body simulations.

Therefore, we can conclude that approximations of RB MD for the fullerene-base nanocars on metal surfaces may account for somewhat faster dynamics as compared to the more accurate all-atom models. The diffusion coefficients estimated in the rigid-body simulations may be 1 or 2 orders larger due to restrictions imposed on the models.

Summary

We have investigated the reliability of RB MD computer simulation methods by analyzing the dynamics of nanocars on gold and silver surfaces. Several problems and issues have been analyzed.

The consequences of utilizing the Nosé–Poincaré thermostat are carefully studied at different conditions. It is found that numerical integration errors do not contribute significantly to determination of the dynamic properties of the system. However, if the trend in errors of measuring quantities continues, it might limit the length of time trajectories that can be used reliably for calculations of properties of the molecules on the surface. The way to improve the accuracy of the RB MD method is to have molecules pre-equilibrated or to neglect early times. In addition, this thermostat is consistent in reproducing canonical ensemble distributions. Thus, the application of the Nosé–Poincaré thermostat is an important part of RB MD simulations that allows this method to be realistic in computing the dynamic properties of the system.

The correlations in the motion of the wheels have been consistently analyzed via several correlation functions. It is found that these correlations are small, but they cannot be neglected.

In addition, we studied the effect of interactions, molecular flexibility, and temperature on the dynamics of single molecules hopping on surfaces. In the case of stronger interactions of the z-car molecule with the surface in comparison with interfragmental forces, these interactions fully determine all dynamic properties. It is also observed that increasing the flexibility raises the overall molecular mobility due to adding the contribution of the rotational motion to the ability to displace the molecule. We have found that the diffusion of nanocars is a temperature-activated process. The activation energy barrier shows a nonmonotonic dependence on the number of rigid fragments used in the simulations. This result is explained by interplay between two factors: increasing the number of fragments leads to more rotations, lowering the effective activation barrier, but individual rotations also allow the fragments to adjust to interact better with the surface, increasing the activation barrier. We also compared dynamical properties of nanocars on the gold and silver surfaces. It is argued that the z-car probably moves faster on Au because on the silver surface the wheels of the molecules are in bent configurations, leading to stronger interactions with the surface and slower overall motion.

Thus, our comprehensive analysis of RB MD computer simulation methods shows that this is a very reliable and highly convenient approach that might be successfully used for understanding the dynamics of nanoscale processes. However, it should be noted also that our calculations indicate that the rigid-body approximation leads to faster dynamics with larger diffusion coefficients (1–2 orders of magnitude) as compared with full atomic MD computer simulations. It will be interesting to extend this approach to describe systems under the effect of external fields (electromagnetic, light, etc.), systems on different surfaces, including nonmetallic, and systems with many moving and interacting species.

Acknowledgment. A.B.K. acknowledges support from the Welch Foundation (Grant C-1559) and from the U.S. National Science Foundation (Grant ECCS-0708765). This work was also supported in part by the Shared University Grid at Rice University and by the Russian Foundation for Basic Research (Grant 09-03-00338). The Russian team acknowledges the facilities of the Supercomputing Complex of the Research Computing Center of M.V. Lomonosov Moscow State University and the SKIF-GRID program for providing computational resources.

References

- (1) Dominguez, Z.; Dang, H.; Strouse, J.; Garcia-Garibay, M. A. *J. Am. Chem. Soc.* **2002**, *124*, 2398.
- (2) Godinez, C. E.; Zepeda, G.; Garcia-Garibay, M. A. *J. Am. Chem. Soc.* **2002**, *124*, 4701.
- (3) Kottas, G. S.; Clarke, L. I.; Horinek, D.; Michl, J. *Chem. Rev.* **2005**, *105*, 1281.
- (4) Balzani, V.; Credi, A.; Venturi, M. *ChemPhysChem* **2008**, *9*, 202.
- (5) van Delden, R. A.; ter Wiel, M. K. J.; Pollard, M. M.; Vicario, J.; Koumura, N.; Feringa, B. L. *Nature* **2005**, *437*, 1337.
- (6) Gimzewski, J. K.; Joachim, C.; Schlittler, R. R.; Langlais, V.; Tang, H.; Johannsen, I. *Science* **1998**, *281*, 531.
- (7) Baber, A. E.; Tierney, H. L.; Sykes, E. C. H. *ACS Nano* **2008**, *2*, 2385.
- (8) Michl, J.; Sykes, E. C. H. *ACS Nano* **2009**, *3*, 1042.
- (9) Vives, G.; Kang, J.; Kelly, K. F.; Tour, J. M. *Org. Lett.* **2009**, *11*, 5602.
- (10) Claytor, K.; Khatua, S.; Guerrero, J.; Tcherniak, A.; Tour, J. M.; Link, S. *J. Chem. Phys.* **2009**, *130*, 164710.
- (11) Khatua, S.; Guerrero, J. M.; Claytor, K.; Vives, G.; Kolomeisky, A. B.; Tour, J. M.; Link, S. *ACS Nano* **2009**, *3*, 351.
- (12) Vives, G.; Tour, J. M. *Acc. Chem. Res.* **2009**, *42*, 473.
- (13) Vives, G.; Tour, J. M. *Tetrahedron Lett.* **2009**, *50*, 1427.
- (14) Sasaki, T.; Guerrero, G.; Leonard, A. D.; Tour, J. M. *Nano Res.* **2008**, *1*, 412.
- (15) Shirai, Y.; Morin, J.-F.; Sasaki, T.; Guerrero, J. M.; Tour, J. M. *Chem. Soc. Rev.* **2006**, *35*, 1043.
- (16) Shirai, Y.; Osgood, A. J.; Zhao, Y.; Yao, Y.; Saudan, L.; Yang, H.; Yu-Hung, C.; Sasaki, T.; Morin, J.-F.; Guerrero, J. M.; Kelly, K. F.; Tour, J. M. *J. Am. Chem. Soc.* **2006**, *128*, 4854.
- (17) Akimov, A. V.; Nemukhin, A. V.; Moskovsky, A. A.; Kolomeisky, A. B.; Tour, J. M. *J. Chem. Theory Comput.* **2008**, *4*, 652.
- (18) Reich, S. *SIAM J. Numer. Anal.* **1996**, *33*, 475.
- (19) Ikeguchi, M. *J. Comput. Chem.* **2004**, *25*, 529.
- (20) Miller III, T. F.; Eleftheriou, M.; Pattnaik, P.; Ndirango, A.; Newns, D.; Martyna, G. J. *J. Chem. Phys.* **2002**, *116*, 8649.
- (21) Omelyan, I. P. *Phys. Rev. E* **1998**, *58*, 1169.
- (22) Omelyan, I. P. *Comput. Phys. Commun.* **1998**, *109*, 171.
- (23) Moskovsky, A. A.; Vanovschi, V. V.; Konyukhov, S. S.; Nemukhin, A. V. *Int. J. Quantum Chem.* **2006**, *106*, 2208.
- (24) Dullweber, A.; Leimkuhler, B.; McLachlan, R. *J. Chem. Phys.* **1997**, *107*, 5840.
- (25) Tierney, H. L.; Baber, A. E.; Sykes, E. C. H.; Akimov, A.; Kolomeisky, A. B. *J. Phys. Chem. C* **2009**, *113*, 10913.
- (26) Horinek, D.; Michl, J. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 14175.
- (27) Horinek, D.; Michl, J. *J. Am. Chem. Soc.* **2003**, *125*, 11900.
- (28) Hou, S.; Sagara, T.; Xu, D.; Kelly, T. R.; Ganz, E. *Nanotechnology* **2003**, *14*, 566.
- (29) Vacek, J.; Michl, J. *Adv. Funct. Mater.* **2007**, *17*, 730.
- (30) Bond, S. D.; Leimkuhler, B. J.; Laird, B. B. *J. Comput. Phys.* **1999**, *151*, 114.
- (31) Nosé, S. *J. Chem. Phys.* **1984**, *81*, 511.
- (32) Nosé, S. *J. Phys. Soc. Jpn.* **2001**, *70*, 75.
- (33) Leimkuhler, B. L.; Sweet, Ch. R. *J. Chem. Phys.* **2004**, *121*, 108.
- (34) Martyna, G. J.; Klein, M. L.; Tuckerman, M. *J. Chem. Phys.* **1992**, *97*, 2635.
- (35) MacKerell, A. D., Jr.; Bashford, D.; Bellott, M.; Dunbrack, R. L., Jr.; 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., III; 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.
- (36) Foloppe, N.; MacKerell, A. D., Jr. *J. Comput. Chem.* **2000**, *21*, 86.
- (37) Rappe, A. K.; Casewit, C. J.; Colwell, K. S.; Goddard, W. A., III; Skiff, W. M. *J. Am. Chem. Soc.* **1992**, *114*, 10024.
- (38) Jalkanen, J. P.; Zerbetto, F. *J. Phys. Chem. B* **2006**, *110*, 5595.
- (39) Wang, L. L.; Cheng, H. P. *Phys. Rev. B* **2004**, *69*, 165417.
- (40) Wang, L. L.; Cheng, H. P. *Phys. Rev. B* **2004**, *69*, 045404.
- (41) Perez-Jimenez, A. J.; Palacios, J. J.; Louis, E.; SanFabian, E.; Verges, J. A. *ChemPhysChem* **2003**, *4*, 388.
- (42) Tzeng, C. T.; Lo, W. S.; Yuh, J. Y.; Chu, R. Y.; Tsuei, K. D. *Phys. Rev. B* **2000**, *61*, 2263.
- (43) Rogero, C.; Pascual, J. I.; Gomez-Herrero, J.; Baro, A. M. *J. Chem. Phys.* **2002**, *116*, 832.
- (44) Yoon, B.; Luedtke, W. D.; Gao, J.; Landman, U. *J. Phys. Chem. B* **2003**, *107*, 5882.

CT100101Y

Water's Contribution to the Energetic Roughness from Peptide Dynamics

Quentin Johnson,[†] Urmi Doshi,[†] Tongye Shen,^{‡,§} and Donald Hamelberg^{*,†}

Department of Chemistry and the Center for Biotechnology and Drug Design, Georgia State University, Atlanta, Georgia 30302-4098, Department of Biochemistry, Cellular & Molecular Biology, University of Tennessee, Knoxville, Tennessee 37996, and Center for Molecular Biophysics, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37830

Received April 5, 2010

Abstract: Water plays a very important role in the dynamics and function of proteins. Apart from protein–protein and protein–water interactions, protein motions are accompanied by the formation and breakage of hydrogen-bonding network of the surrounding water molecules. This ordering and reordering of water also adds to the underlying roughness of the energy landscape of proteins and thereby alters their dynamics. Here, we extract the contribution of water to the ruggedness (in terms of an energy scale ε) of the energy landscape from molecular dynamics simulations of a peptide substrate analogue of prolyl *cis*–*trans* isomerases. In order to do so, we develop and implement a model based on the position space analog of the Ornstein–Uhlenbeck process and Zwanzig's theory of diffusion on a rough potential. This allows us to also probe an important property of the widely used atomistic simulation water models that directly affects the dynamics of biomolecular systems and highlights the importance of the choice of the water model in studying protein dynamics. We show that water contributes an additional roughness to the energy landscape. At lower temperatures this roughness, which becomes comparable to $k_B T$, can considerably slow down protein dynamics. These results also have much broader implications for the function of some classes of enzymes, since the landscape topology of their substrates may change upon moving from an aqueous environment into the binding site.

Introduction

The hyperdimensional energy landscape of biological macromolecules has a very complex topology representing self-assembly of their three-dimensional structures from unfolded states, their interactions with other partners, and conformational transitions for their function.^{1,2} To add to the complexity, the surface of the landscape is not smooth but rather jagged with numerous local minima separated by energy

barriers of different heights. These minima represent a vast number of conformational substates, each specified by the configuration of the biomolecule including its hydration shell (i.e., water molecules immediately surrounding the biomolecule) as well as the internal water molecules. Water thus constitutes an integral part of biomolecular structure.^{1,3,4} As the biomolecule undergoes conformational switching between these substates, energy barriers arise due to the formation and breakage of unfavorable interactions within the biomolecule, between the biomolecule and water, and between water molecules. The unevenness of the energy landscape can lead to kinetic traps if it is comparable or much greater than $k_B T$, where k_B is the Boltzmann constant and T is the temperature. To highlight the nature of the inherent roughness of the energy landscape of biomolecules, theories have been developed^{5–7} and for proteins, roughness has been estimated

* Corresponding author phone: (404)413-5564; fax: (404)413-5505; e-mail: dhamelberg@gsu.edu. Corresponding author address: Department of Chemistry, Georgia State University, Atlanta, GA 30302-4098.

[†] Georgia State University.

[‡] University of Tennessee.

[§] Oak Ridge National Laboratory.

experimentally.^{6,8,9} From these studies the characteristic scale for roughness is estimated to fall in the range of few (i.e., $\sim 2-5$) $k_B T$ at 25 °C.

The potential energy landscape of water is also rough with many minima, each typifying a different configuration of the hydrogen-bonding network. Therefore, water by itself has slow dynamics as the hydrogen-bonding network rearranges by breaking and forming hydrogen bonds.^{10,11} The energetic component of hydrogen bond network rearrangements has been estimated to be between 0.8 and 1.5 kcal/mol using X-ray absorption spectroscopy.^{12,13} This value represents the average thermal energy required to distort a hydrogen bond or to rearrange or change the fully coordinated configuration of water to a configuration with a broken hydrogen bond to the donor. However, the quality of the data and the interpretation of the results have been questioned.¹⁴ The exact structure of liquid water and the average thermal energy associated with hydrogen bond rearrangements are still unresolved and involve areas of active research.¹⁵⁻¹⁹ Nonetheless, reorientation of hydrogen bonds between water molecules around proteins will undoubtedly have an effect on the dynamics of proteins and will show up as an energetic component on the overall protein energy landscape.

Biomolecular motions, especially those that bring about conversions between different conformational substates and do not involve forming or breaking of any covalent bonds, are coupled to solvent fluctuations.^{20,21} The role of water on protein dynamics has been studied extensively.²⁰⁻²³ It has been suggested from spectroscopic studies that large-scale protein motions such as folding/unfolding or conformational changes associated with opening/closing of protein channels or accommodation of ligands are dampened by the fluctuations in the bulk solvent and therefore depend on solvent viscosity.²¹ The water molecules in the hydration shell form a short-lived clathrate-like structure around the protein molecules. As proteins undergo constant thermal motions and conformational changes, the network of hydrogen bonds in the hydration shell rearranges. Breaking and reforming of these hydrogen bonds contribute to the overall roughness on the energy landscape of the protein and can restrict the motion of proteins, especially at lower temperatures. Besides, the effect of solvent viscosity, which is manifested as frictional drag and at the microscopic level as the dynamic network of hydrogen bonds in bulk water, also partly adds to the energetic roughness and impedes protein dynamics. In order to get a better insight into the function and dynamics of proteins, it is necessary to understand the nature of the roughness and calculate the magnitude of the different contributions to it. The magnitude of the roughness due to hydration on the energy landscape is not well characterized, and its effects on protein dynamics are less understood. A full understanding of this aspect of protein dynamics has broad implications, such as the dynamic effects of desolvated molecules relative to that of solvated molecules, the low-temperature behavior of solvated biomolecules, and the fundamental nature of water hydrogen bond network.

In the peptidyl prolyl *cis-trans* isomerases (PPIases) class of enzymes wherein the active site is very hydrophobic, the

effect of moving the substrate from aqueous solvent to the active site of PPIases has been suggested to be a possible contributing factor to the catalytic activity.²³⁻²⁵ PPIases catalyze the *cis-trans* isomerization of the peptide ω -bond preceding prolyl of protein backbone, and they function without any bond forming or breaking during the catalytic process.²³ *Cis-trans* isomerization of the prolyl ω -bond is the switching mechanism in several signaling pathways²⁶ and is important for protein folding.^{27,28} The role of aqueous solvent, or the lack thereof, in PPIase catalytic activity is not well understood. Therefore, in the present work we investigate the effects of solvent on the dynamics of a peptide substrate analogue (Ace-Ala-Pro-Nme), focusing only on one torsional degree of freedom, i.e. the peptide bond dihedral (ω) preceding Pro. The ω -bond angle is a good reaction coordinate for describing the *cis-trans* isomerization.²⁹ We present here a novel approach to quantitatively capture the energetic effect of water on the roughness along the prolyl ω -bond angle. In the peptide substrate analogue, Ace-Ala-Pro-Nme, we observe the fluctuations of the ω dihedral preceding the Pro residue in the *trans* basin in all-atom molecular dynamics (MD) simulations with explicit water. We develop a model by describing the Brownian motion in a harmonic well as a position space analog of the Ornstein-Uhlenbeck process.³⁰ We then calculate the diffusion coefficients of the ω dihedral angle on an effective 1-D harmonic potential at different temperatures in the presence of each of the commonly used water models. Using the temperature dependence of these effective diffusion coefficients and the expression given by Zwanzig⁵ that links the effective diffusion coefficient to roughness, we tease out the magnitude of the (energetic) roughness contributed by various water models. Additionally, we learn the extent to which the roughness provided by water influences the energy landscape and hence the peptide dynamics along the ω -bond angle degree of freedom. Since atomistic simulations are routinely used to extract dynamic information of biomolecules, knowing the effects of solvent on protein dynamics on a quantitative level will allow us to make a more informed choice of the water model to be used in MD studies, i.e. it will help in reproducing experimentally measured quantities of biomolecules better as well as in force field parametrization.

Results

Dynamics of the Protein Peptide ω -Bond Angle. During the course of the MD simulations of Ace-Ala-Pro-Nme (Figure 1, left panel) in explicit water using commonly used atomistic solvent models, we monitored the dynamics of the ω -bond angle (CA-C-N'-CA'). Shown in Figure 1 (right panel) are the fluctuations of the ω -dihedral in the *trans* basin, i.e. around 180°. For biomolecules, in general, the decay of the autocorrelation function of the velocity along a degree of freedom has a characteristic time scale that is much shorter than that of the displacement of that degree of freedom. Since this is also true for the ω -bond angle, we can describe the actual complicated motion of the peptide along the ω -bond angle as diffusive motion on an effective one-dimensional (1D) energy profile, $U(\omega)$. Such diffusive motion of the peptide on an

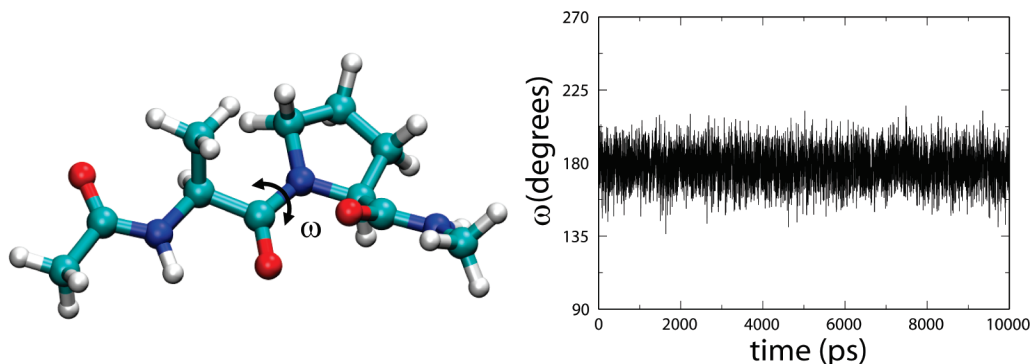


Figure 1. (left) Structure of Ace-Ala-Pro-Nme and (right) the fluctuations of the ω -bond angle between Ala and Pro during the simulation at 300 K in SPC/E water.

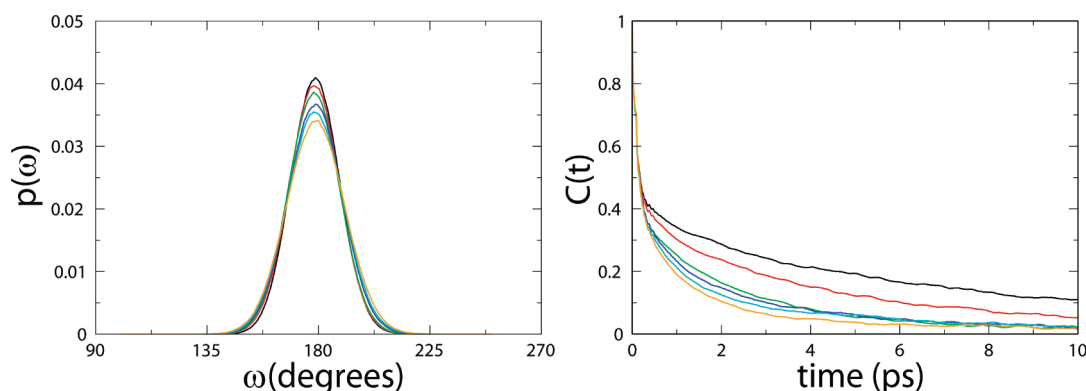


Figure 2. (left) The distribution of the peptide ω -bond angle between Ala and Pro in Ace-Ala-Pro-Nme at six different temperatures (275 K (black), 300 K (red), 325 K (green), 350 K (blue), 375 K (cyan), and 400 K (orange)) in SPC/E water. (right) Autocorrelation function of ω at the same six temperatures as in the left panel.

effective 1D energy profile is generally described by the Smoluchowski equation

$$\frac{\partial p(\omega, t)}{\partial t} = D \frac{\partial}{\partial \omega} \left[\frac{\partial p(\omega, t)}{\partial \omega} + \frac{p(\omega, t)}{k_B T} \frac{\partial U(\omega)}{\partial \omega} \right] \quad (1)$$

where $p(\omega, t)$ is the time-dependent probability distribution of the peptide ω -bond angle, and the diffusion coefficient D is assumed to be independent of ω .

As can be seen from Figure 2, the probability distribution of the ω -bond angle in the *trans* basin is approximately Gaussian. Therefore, the effective 1D potential landscape, $U(\omega)$, of the motion of the peptide along the ω -bond angle in the *trans* basin can be approximated as a harmonic potential $U(\omega) = (K)/(2)(\omega - \gamma)^2$, where K is the effective spring constant and $\gamma \approx 180^\circ$. Furthermore, the Brownian motion in a harmonic potential described as a position space analog of the Ornstein–Uhlenbeck (OU) process³⁰ has been studied extensively, and the autocorrelation function of ω is given by

$$C(t) = \frac{\langle \omega(0)\omega(t) \rangle}{\langle \omega^2 \rangle} = \exp(-tDK/k_B T) \quad (2)$$

where

$$\langle \omega^2 \rangle = k_B T / K \quad (3)$$

The autocorrelation functions of ω at six different temperatures (275, 300, 325, 350, 375, and 400 K) calculated from the simulation data in the SPC/E water model are also shown in Figure 2 (right panel). By fitting the tail of the autocorrelation functions in Figure 2 using eqs 2 and 3 to a single exponential, we obtained D , the diffusion coefficient, of the displacement of the ω -bond angle on the effective 1D harmonic well at different temperatures.

Roughness Contributed by Different Water Models. By analytically solving the Smoluchowski equation (eq 1), Zwanzig⁵ has shown previously that the diffusion coefficient, D , on an effective 1D landscape can be related to the underlying energetic roughness, ϵ , by

$$D = D_0 \exp[-(\epsilon/k_B T)^\theta] \quad (4)$$

where D_0 is the diffusion coefficient on the smooth potential energy surface. Subsequently, it was shown that for a protein system in an effective 1D energy profile $\theta = 2$.^{31,32} The quadratic dependence in eq 4 implies that the energetic roughness is random and has a Gaussian distribution. If $\theta = 1$, then the roughness is uniform and evenly distributed.⁵

From the diffusion coefficients, D , derived from fitting the autocorrelation function to an OU process at different temperatures, we obtained a plot of eq 4, as shown in Figure 3 for the simulations in the different water models³³ (i.e., TIP3P, SPC/E, TIP4P-Ew, and TIP5P) and also in vacuum.

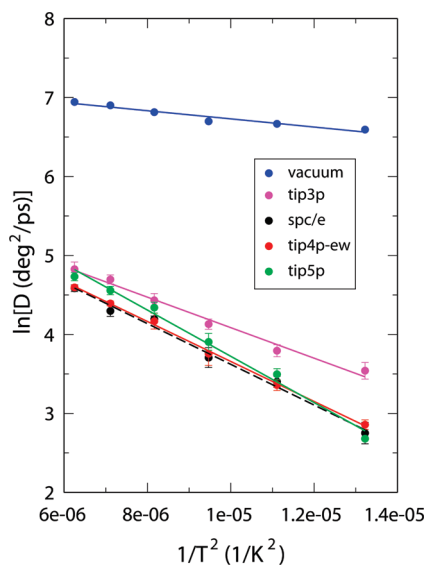


Figure 3. Temperature dependence of the effective diffusion coefficients of the ω -bond angle on the 1D harmonic well from simulations in four different water models and in vacuum. Linear fits using eq 4 with $\theta = 2$.

Table 1. Comparison of Energetic Roughness, Self-Diffusion Coefficients, and Structures of Different Water Models along with Intrinsic Roughness of the Peptide in Vacuum and Experimental Self-Diffusion Coefficient of Water

Water model	Roughness, ϵ (kcal/mol)	Self-diffusion coefficient ($\times 10^5$ cm ² /s) at $\sim 25^\circ\text{C}$ and 1 atm	Structure of water model (A ball and stick model with partial charges on each charge center)
Vacuum	0.45	-	
TIP3P ³⁴	0.87	5.06 ³⁵ 5.65 ³⁶	
SPC/E ³⁷	1.01	2.49 ³⁵ 2.76 ³⁶	
TIP4P-Ew ^{38,39}	1.00	2.4 ³⁸	
TIP5P ⁴⁰	1.08	2.62 ³⁵	
Experimental	-	2.23 ⁴¹ 2.3 ^{42,43}	

For each water model the data fitted very well with $\theta = 2$. By increasing the temperature range and carrying out simulations in TIP4P-EW at 600 K we further confirmed that $\theta = 2$ gives a better fit than $\theta = 1$. The plot of $\ln(D)$ versus $1/T^2$ ($R^2=0.98$) exhibited a higher correlation coefficient than $\ln(D)$ versus $1/T$ ($R^2=0.93$). Using $\theta = 2$ we calculated the roughness contributed by the different water models as well as the peptide's intrinsic roughness along the ω -bond angle in vacuum (the baseline) that are shown in Table 1. From Figure 3, it can be seen that the dynamics of the backbone ω angle is noticeably different in TIP3P

than in SPC/E, TIP4P-Ew, and TIP5P, especially at lower temperature, i.e. the diffusion coefficients in TIP3P are higher as compared to those in other water models. The smaller contribution provided by TIP3P to the roughness results in faster dynamics at lower temperature. For the temperature range considered in this study, it is interesting to note the difference in backbone dynamics in the presence of different water models even at ambient to high temperatures.

If one considers the roughness of Ace-Ala-Pro-Nme in vacuum as the baseline, then TIP5P contributes ~ 0.63 kcal/mol of additional roughness to the energy landscape. TIP4P-EW and SPC/E contribute about ~ 0.56 kcal/mol. The roughness contributed by TIP3P (~ 0.42 kcal/mol) is the smallest. Therefore, SPC/E, TIP4P-EW, and TIP5P are slightly “rougher” than TIP3P. It is also important to note that the self-diffusion coefficient of TIP3P is about twice that of SPC/E, TIP4P-EW, and TIP5P (see Table 1), and the differences in energetic roughness can be partly attributed to that as well.

What Is the Main Contributing Factor to the Energetic Roughness by Water? By looking at the partial charges on the water models with three charged centers (SPC/E, TIP3P, and TIP4P-EW), one can see that the absolute magnitude of the partial charges is larger in SPC/E and TIP4P-EW than in TIP3P, and the increase correlates with a slight increase in the energetic roughness. This observation therefore raises the possibility that the majority of the contribution to the energetic roughness can be due to hydrogen bonding, which is electrostatic in nature in the classical definition of the water models. Alternatively, one can argue that the roughness may have an artificial component due to the Langevin thermostat and the random noise associated with it. However, it is important to note that the roughness is distinctly different for the different water models under similar conditions and using the same thermostat. Nonetheless, in order to test the effect of the Langevin thermostat on the energetic roughness, we doubled the collision frequency to 20 ps^{-1} and repeated the simulations for TIP4P-EW water. The diffusion coefficients obtained from these simulations at various temperatures are shown in Figure 4 (magenta). We clearly see, as expected at higher friction, that overall the dynamics is slightly slower compared to that obtained using a collision frequency of 10 ps^{-1} (Figure 4; red line). However, the slope of the line and hence the roughness is almost identical to simulations with a smaller collision frequency (Figure 4, compare red and magenta lines). Thus, changing the collision frequency or friction only affects D_0 in eq 4, as has also been observed earlier.³²

Consequently, we hypothesize that the main source of energetic roughness is due to the forming and breaking of hydrogen bonding interactions between the water molecules, since the other nonbonded van der Waals interactions are far weaker. As noted above, the description of the hydrogen bonding interaction in the current classical model is purely electrostatic. Therefore, in order to test this hypothesis, we repeated the simulations of the peptide in a modified version of TIP4P-EW water, by zeroing all the partial charges only on the water molecules, thus eliminating the electrostatic interactions between the water molecules. The simulations

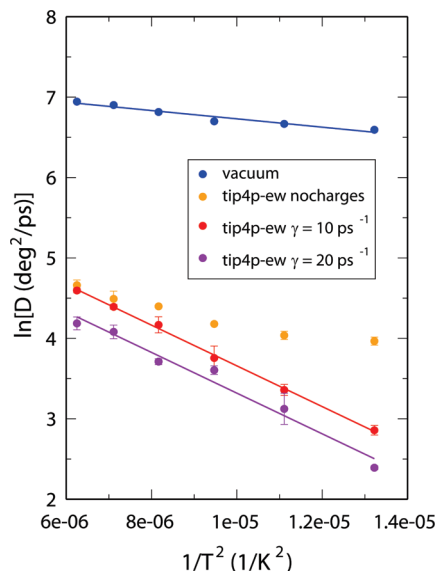


Figure 4. A plot of $\ln D$ versus $1/T^2$ for the simulations in vacuum and in TIP4P-EW when the collision frequency for the Langevin thermostat was set to 10 and 20 ps^{-1} and when the partial charges were set to zero. Linear fits using eq 4 with $\theta = 2$.

were carried out with the same box sizes as those of the simulations with full partial charges at the different temperatures, in order to capture only the effect of eliminating the electrostatic interactions. For this model system of TIP4P-EW water with no partial charges, we clearly see that the slope of the line and hence the roughness also decreases considerably (Figure 4, orange) and is now comparable to that in vacuum. However, the data for TIP4P-EW with zero partial charges fitted eq 4 a little better with $\theta = 1$ ($R^2=0.96$) than with $\theta = 2$ ($R^2=0.94$), implying that without the electrostatic interactions the roughness is more uniform and evenly distributed. This change in the nature of the roughness can be attributed to the fact that van der Waals interactions are very short ranged and due only to the oxygen, since the radius of hydrogen for these water models is zero. On the other hand, electrostatic interactions are longer-ranged, and a slight change can have implications far away, adding to the randomness of this interaction.

Discussion and Concluding Remarks

We have investigated the effects of hydration on protein backbone motions and estimated the energetic roughness contributed by the most widely used water models on a quantitative level from all-atom MD simulations. We find that TIP3P water provides the least roughness resulting in the fastest dynamics of the ω -bond angle compared to that in other water models. This is not surprising since TIP3P has the highest translational diffusion coefficient ($\sim 5 \times 10^{-5} \text{ cm}^2/\text{s}$), which is more than double the self-diffusion coefficient of water ($\sim 2.3 \times 10^{-5} \text{ cm}^2/\text{s}$) estimated from experiments.⁴² Similar behavior of the TIP3P water model has been observed in estimating the rotational diffusion of folded proteins using different atomistic simulation water models.⁴⁴ Furthermore, we note that all the other commonly used water models (SPC/E,

TIP4P-Ew, TIP5P) considered in this study also have self-diffusion coefficients a little higher than that of water from experimental estimates. This means that protein dynamics with these water potentials will be slightly faster than in real water, and the energetic roughness provided by water may be slightly higher in reality than what it may appear from atomistic simulations. Hence real water may be slightly “rougher” than most of the simulation water models.

Interestingly, we also find that for the range of temperatures studied here (275 K–400 K) the dynamic behavior of protein backbone varies depending on the choice of the water model with the differences becoming more pronounced as temperatures are lowered. However, in contrast, a previous study of extremely low-temperature (20–300 K) behaviors of myoglobin solvated in TIP3P, TIP4P, and TIP5P has concluded that dynamic properties of proteins characterized by average mean-square displacements and time-averaged structures are similar irrespective of the choice of any TIP model.⁴⁵

This work presents a very important and an additional property of water that can be taken into account while optimizing water potentials and, if need be, to reproduce experimental results. Such improved water potentials will help in more accurate simulations of protein motions. Our work also attempts to explain the underlying factors responsible for the roughness originating from the solvent. As depicted in Figure 5, the network of hydrogen bonds formed by water molecules around proteins constantly rearranges as the protein changes conformation. The water molecules immediately around the peptide form a transient, cagelike structure that is held together by the network of hydrogen bonds (Figure 5). As the conformation of the peptide changes, the hydrogen bonds break and reform to alter the network of hydrogen bonds. Each arrangement of the network of water molecules is an energetic substate of water and manifests itself as additional roughness on the underlying energy surface of proteins. The energetic roughness can have several implications for protein chemistry and motions. For example, when $\theta = 2$ in eq 4, the effect of the roughness on the dynamics of the protein can become very pronounced at low temperatures and restricts the motions of proteins. At higher temperature, when $k_B T$ is much higher than the roughness (which is $\sim 1 \text{ kcal/mol}$), the effect of water molecules on the dynamics will be predominately due to the collision of the molecules and less due to the roughness of water.

Additionally, proteins that function as enzymes usually provide an environment in a cavity or binding site that is usually devoid of any appreciable amount of solvent. The effect of moving the substrate from the aqueous medium to the active site on the catalytic process is still not fully understood. However, this effect may only be limited to certain classes of enzymes,⁴⁶ like PPIases. From the above results, it can be suggested that the change in environment of the substrate can change the frictional drag as well as the topological features of the energy landscape of the substrate and thus “pave” the surface along the reaction coordinate for the conformational transition to occur. This

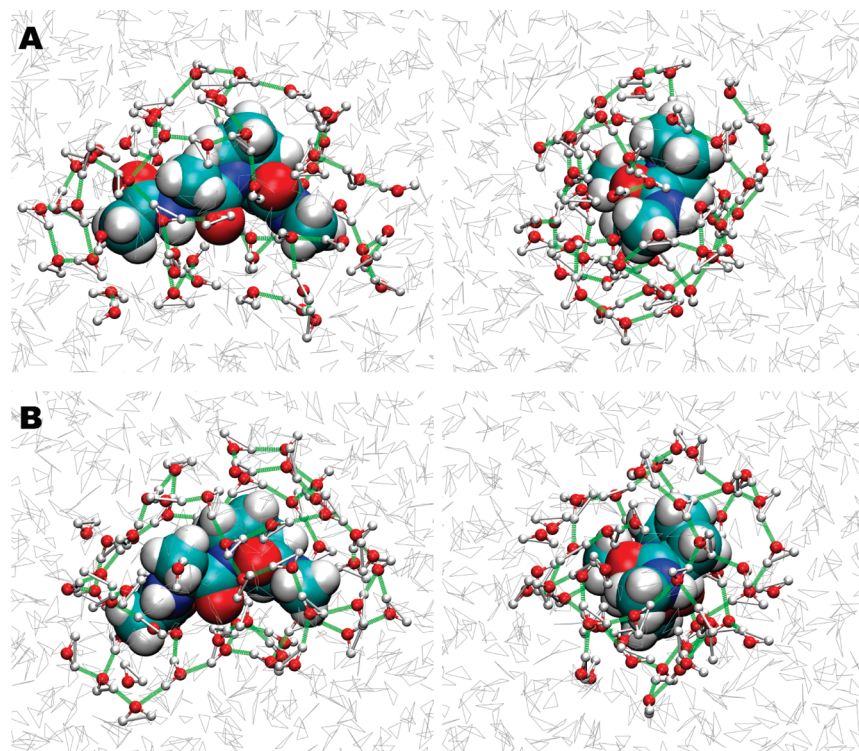


Figure 5. Hydrogen bonding network of the surrounding water molecules when the peptide Ace-Ala-Pro-Nme is in two different configurations (A and B). Each panel shows two different views of a configuration. The immediate water molecules surrounding this peptide are shown as balls and sticks (red and white), and the hydrogen bonds between the water molecules are shown as green. The remaining water molecules are shown as gray lines.

change in the topology of the landscape will most likely manifest itself by altering the kinetic pre-exponential factor. However, the dominating effect on the catalysis will still come from preorganization of the active site⁴⁷ and transition state stabilization, as shown for PPIases⁴⁸ or other energetic effects that are part of the exponential component of the kinetic rate equation. An indirect evidence for this effect is observed from the rate of *cis-trans* isomerization in different solvents. It can be clearly seen experimentally that the rate of isomerization is directly correlated with the solvents' self-diffusion coefficient,²⁴ which is inversely proportional to solvent friction.

In the current study, we have extracted information on peptide motions along one of the backbone torsional angles, i.e. ω -bond angle. In order to do so, we have used a simplified OU model by taking advantage of a good overdamped harmonic oscillator approximation of the dynamics. For such approximation to work, it is required that the degree of freedom in question has a Gaussian distribution. It will be interesting to investigate whether the same model can be applied to study peptide dynamics along other degrees of freedom and see how solvent alters those motions. It is possible that modified dynamic models may be required to reflect the characteristics of the free energy profiles along other degrees of freedom. A comparative study of water's contribution to the energetic roughness along various degrees of freedom may provide a clearer and more quantitative picture of the energy landscape of peptides. Moreover, such studies will also provide clues on how solvent affects the dynamics of

folded proteins and the topological features of the potential energy landscape in the native well.

Computational Methods

Using the model peptide Ace-Ala-Pro-Nme we carried out a series of MD simulations in four widely used explicit simulation water models: TIP3P,³⁴ SPC/E,³⁷ TIP4P-EW,^{38,39} and TIP5P.⁴⁰ For each water model, the simulations were performed at 275 K, 300 K, 325 K, 350 K, 375 K, and 400 K using the pmemd module in the AMBER 10 suite of programs⁴⁴ and the modified version⁴⁹ of the Cornell et al.⁵⁰ force field. The system was equilibrated at the set temperature and a constant pressure of 1 bar in a periodic cubic box with the edges of the box at least 10 Å away from the peptide. The temperature was regulated using the Langevin thermostat with a collision frequency of 10 ps⁻¹. Particle mesh Ewald⁵¹ method was used to treat long-range interactions and all bonds involving hydrogen atoms were constrained by applying the SHAKE⁵² algorithm. Three independent MD runs were carried out for 10 ns each at constant temperature and volume (NVT ensemble) using an integration time step of 2 fs for every combination of water model and temperature.

Acknowledgment. This work was supported in part by the National Science Foundation CAREER MCB-0953061 (D.H.) and Georgia Cancer Coalition (D.H.). This work was also supported by Georgia State's IBM System p5 super-computer, acquired through a partnership of the Southeastern Universities Research Association and IBM supporting the SURagrid initiative.

References

- (1) Frauenfelder, H.; Sligar, S. G.; Wolynes, P. G. *Science* **1991**, *254*, 1598.
- (2) Wales, D. J. *Energy landscapes*; Cambridge University Press: Cambridge, UK, New York, 2003.
- (3) Frauenfelder, H.; Parak, F.; Young, R. D. *Annu. Rev. Biophys. Chem.* **1988**, *17*, 451.
- (4) Chaplin, M. *Nat. Rev. Mol. Cell Biol.* **2006**, *7*, 861.
- (5) Zwanzig, R. *Proc. Natl. Acad. Sci. U. S. A.* **1988**, *85*, 2029.
- (6) Hyeon, C.; Thirumalai, D. *Proc. Natl. Acad. Sci. U. S. A.* **2003**, *100*, 10249.
- (7) Sagnella, D. E.; Straub, J. E.; Thirumalai, D. *J. Chem. Phys.* **2000**, *113*, 7702.
- (8) Lapidus, L. J.; Eaton, W. A.; Hofrichter, J. *Proc. Natl. Acad. Sci. U. S. A.* **2000**, *97*, 7220.
- (9) Nevo, R.; Brumfeld, V.; Kapon, R.; Hinterdorfer, P.; Reich, Z. *EMBO Rep.* **2005**, *6*, 482.
- (10) Ohmine, I.; Saito, S. *Acc. Chem. Res.* **1999**, *32*, 741.
- (11) Speedy, R. J.; Madura, J. D.; Jorgensen, W. L. *J. Phys. Chem.* **1987**, *91*, 909.
- (12) Smith, J. D.; Cappa, C. D.; Wilson, K. R.; Messer, B. M.; Cohen, R. C.; Saykally, R. J. *Science* **2004**, *306*, 851.
- (13) Smith, J. D.; Cappa, C. D.; Messer, B. M.; Cohen, R. C.; Saykally, R. J. *Science* **2005**, *308*, 793B.
- (14) Nilsson, A.; Wernet, P.; Nordlund, D.; Bergmann, U.; Cavalleri, M.; Odellius, M.; Ogasawara, H.; Naslund, L. A.; Hirsch, T. K.; Glatzel, P.; Pettersson, L. G. M. *Science* **2005**, *308*, 793A.
- (15) Chumaevskii, N. A.; Rodnikova, M. N. *J. Mol. Liq.* **2003**, *106*, 167.
- (16) Head-Gordon, T.; Johnson, M. E. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103*, 7973.
- (17) Markovitch, O.; Agmon, N. *Mol. Phys.* **2008**, *106*, 485.
- (18) Myneni, S.; Luo, Y.; Naslund, L. A.; Cavalleri, M.; Ojamae, L.; Ogasawara, H.; Pelmenchikov, A.; Wernet, P.; Vaterlein, P.; Heske, C.; Hussain, Z.; Pettersson, L. G. M.; Nilsson, A. *J. Phys.: Condens. Matter* **2002**, *14*, L213.
- (19) Wernet, P.; Nordlund, D.; Bergmann, U.; Cavalleri, M.; Odellius, M.; Ogasawara, H.; Naslund, L. A.; Hirsch, T. K.; Ojamae, L.; Glatzel, P.; Pettersson, L. G. M.; Nilsson, A. *Science* **2004**, *304*, 995.
- (20) Fenimore, P. W.; Frauenfelder, H.; McMahon, B. H.; Parak, F. G. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99*, 16047.
- (21) Frauenfelder, H.; Chen, G.; Berendzen, J.; Fenimore, P. W.; Jansson, H.; McMahon, B. H.; Stroe, I. R.; Swenson, J.; Young, R. D. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 5129.
- (22) Dill, K. A. *Biochemistry* **1990**, *29*, 7133.
- (23) Fanghanel, J.; Fischer, G. *Front. Biosci.* **2004**, *9*, 3453.
- (24) Eberhardt, E. S.; Loh, S. N.; Hinck, A. P.; Raines, R. T. *J. Am. Chem. Soc.* **1992**, *114*, 5437.
- (25) Ikura, T.; Kinoshita, K.; Ito, N. *Protein Eng. Des. Sel.* **2008**, *21*, 83.
- (26) Lu, K. P.; Finn, G.; Lee, T. H.; Nicholson, L. K. *Nat. Chem. Biol.* **2007**, *3*, 619.
- (27) Brandts, J. F.; Halvorson, H. R.; Brennan, M. *Biochemistry* **1975**, *14*, 4953.
- (28) Wedemeyer, W. J.; Welker, E.; Scheraga, H. A. *Biochemistry* **2002**, *41*, 14637.
- (29) Doshi, U.; Hamelberg, D. *J. Phys. Chem. B* **2009**, *113*, 16590.
- (30) Wang, M. C.; Uhlenbeck, G. E. *Rev. Mod. Phys.* **1945**, *17*, 20.
- (31) Hamelberg, D.; Shen, T.; Andrew McCammon, J. *J. Chem. Phys.* **2005**, *122*, 241103.
- (32) Hamelberg, D.; Shen, T.; McCammon, J. A. *J. Chem. Phys.* **2006**, *125*, 094905.
- (33) Guillot, B. *J. Mol. Liq.* **2002**, *101*, 219.
- (34) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926.
- (35) Mahoney, M. W.; Jorgensen, W. L. *J. Chem. Phys.* **2001**, *114*, 363.
- (36) Mark, P.; Nilsson, L. *J. Phys. Chem. A* **2001**, *105*, 9954.
- (37) Berendsen, H. J. C.; Grigera, J. R.; Straatsma, T. P. *J. Phys. Chem.* **1987**, *91*, 6269.
- (38) Horn, H. W.; Swope, W. C.; Pitera, J. W.; Madura, J. D.; Dick, T. J.; Hura, G. L.; Head-Gordon, T. *J. Chem. Phys.* **2004**, *120*, 9665.
- (39) Jorgensen, W. L.; Madura, J. D. *Mol. Phys.* **1985**, *56*, 1381.
- (40) Mahoney, M. W.; Jorgensen, W. L. *J. Chem. Phys.* **2000**, *112*, 8910.
- (41) Gillen, K. T.; Douglas, D. C.; Hoch, M. J. R. *J. Chem. Phys.* **1972**, *57*, 5117.
- (42) Holz, M.; Heil, S. R.; Sacco, A. *Phys. Chem. Chem. Phys.* **2000**, *2*, 4740.
- (43) Price, W. S.; Ide, H.; Arata, Y. *J. Phys. Chem. A* **1999**, *103*, 448.
- (44) Case, D. A.; Darden, T. A.; Cheatham, T. E., III; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Crowley, M.; Walker, R. C.; Zhang, W.; Merz, K. M.; Wang, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Kolossváry, I.; Wong, K. F.; Paesani, F.; Vanicek, J.; Wu, X.; Brozell, S. R.; Steinbrecher, T.; Gohlke, H.; Yang, L.; Tan, C.; Mongan, J.; Hornak, V.; Cui, G.; Matthews, D. H.; Seetin, M. G.; Sagui, C.; Babin, V.; Kollman, P. A. *AMBER 10*; University of California: San Francisco, 2008.
- (45) Glass, D. C.; Krishnan, M.; Nutt, D. R.; Smith, J. C. *J. Chem. Theory Comput.* **2010**, *6*, 1390.
- (46) Warshel, A.; Aqvist, J.; Creighton, S. *Proc. Natl. Acad. Sci. U. S. A.* **1989**, *86*, 5820.
- (47) Warshel, A. *Proc. Natl. Acad. Sci. U. S. A.* **1978**, *75*, 5250.
- (48) Hamelberg, D.; McCammon, J. A. *J. Am. Chem. Soc.* **2009**, *131*, 147.
- (49) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. *Proteins* **2006**, *65*, 712.
- (50) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. *J. Am. Chem. Soc.* **1995**, *117*, 5179.
- (51) Darden, T.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993**, *98*, 10089.
- (52) Ryckaert, J.; Cicotti, G.; Berendsen, H. *J. Comput. Phys.* **1977**, *23*, 327.

JCTC

Journal of Chemical Theory and Computation

Enhanced Conformational Sampling in Molecular Dynamics Simulations of Solvated Peptides: Fragment-Based Local Elevation Umbrella Sampling

Halvor S. Hansen,[†] Xavier Daura,^{‡,§} and Philippe H. Hünenberger^{*,†}

Laboratorium für Physikalische Chemie, ETH Zürich, CH-8093 Zürich, Switzerland,
Institute of Biotechnology and Biomedicine, Universitat Autònoma de Barcelona,
E-08193 Bellaterra (Barcelona), Spain, and Catalan Institution for Research and
Advanced Studies (ICREA), E-08010 Barcelona, Spain

Received June 5, 2010

Abstract: A new method, fragment-based local elevation umbrella sampling (FB-LEUS), is proposed to enhance the conformational sampling in explicit-solvent molecular dynamics (MD) simulations of solvated polymers. The method is derived from the local elevation umbrella sampling (LEUS) method [Hansen and Hünenberger, *J. Comput. Chem.* **2010**, *31*, 1–23], which combines the local elevation (LE) conformational searching and the umbrella sampling (US) conformational sampling approaches into a single scheme. In LEUS, an initial (relatively short) LE build-up (searching) phase is used to construct an optimized (grid-based) biasing potential within a subspace of conformationally relevant degrees of freedom, which is then frozen and used in a (comparatively longer) US sampling phase. This combination dramatically enhances the sampling power of MD simulations but, due to computational and memory costs, is only applicable to relevant subspaces of low dimensionalities. As an attempt to expand the scope of the LEUS approach to solvated polymers with more than a few relevant degrees of freedom, the FB-LEUS scheme involves an US sampling phase that relies on a superposition of low-dimensionality biasing potentials optimized using LEUS at the fragment level. The feasibility of this approach is tested using polyalanine (poly-Ala) and polyvaline (poly-Val) oligopeptides. Two-dimensional biasing potentials are preoptimized at the monopeptide level, and subsequently applied to all dihedral-angle pairs within oligopeptides of 4, 6, 8, or 10 residues. Two types of fragment-based biasing potentials are distinguished: (i) the basin-filling (BF) potentials act so as to “fill” free-energy basins up to a prescribed free-energy level above the global minimum; (ii) the valley-digging (VD) potentials act so as to “dig” valleys between the (four) free-energy minima of the two-dimensional maps, preserving barriers (relative to linearly interpolated free-energy changes) of a prescribed magnitude. The application of these biasing potentials may lead to an impressive enhancement of the searching power (volume of conformational space visited in a given amount of simulation time). However, this increase is largely offset by a deterioration of the statistical efficiency (representativeness of the biased ensemble in terms of the conformational distribution appropriate for the physical ensemble). As a result, it appears difficult to engineer FB-LEUS schemes representing a significant improvement over plain MD, at least for the systems considered here.

1. Introduction

Classical atomistic simulations, in particular molecular dynamics (MD), represent nowadays a powerful tool comple-

mentary to experiment for investigating the properties of atomic and molecular systems relevant in physics, chemistry, and biology.^{1–4}

The success of these methods in the context of condensed-phase systems results in particular from a favorable trade-off between model resolution and computational cost. On the one hand, classical atomistic models, although they represent an approximation to quantum mechanics, can still provide a realistic description of many molecular systems

* Corresponding author. Phone: +41 44 632 5503. Fax: +41 44 632 1039. E-mail: phil@igc.phys.chem.ethz.ch.

[†] ETH Zürich.

[‡] Universitat Autònoma de Barcelona.

[§] Catalan Institution for Research and Advanced Studies (ICREA).

at spatial and temporal resolutions on the order of 0.1 nm and 1 fs, respectively. On the other hand, their computational cost remains tractable at present for system sizes and time scales on the order of 10 nm and 100 ns, respectively. These scales are sufficient to enable in many cases (i) an appropriate description of bulk-like solvation (discrete solvent molecules, sufficiently large solvation range), (ii) a reliable calculation of thermodynamic properties via statistical mechanics (converged ensemble averages and free energies), and (iii) a direct comparison with experimental data (structural, thermodynamic, transport, and dynamic observables measured on similar spatial and temporal scales).

In practice, however, the results of atomistic simulations are still affected by four main sources of error, originating from (i) the classical atomistic approximation^{5–8} (neglect or mean-field treatment of electronic and quantum effects), (ii) the approximate force-field representation of interatomic interactions^{2–4,9} (potential energy function with simplified functional form and empirical parameters, various parameter transferability and combination assumptions), (iii) the presence of finite-size and surface effects^{10,11} (related to the still microscopic size of the simulated systems), and (iv) the insufficient conformational sampling^{9,12–15} (related to the still very short time scale of the simulations). The reduction of the last type of errors can be viewed as a first-priority target in the improvement of simulation methodologies, because insufficient sampling results in errors that are predominantly nonsystematic, while the three other types of errors are systematic.

Among the applications of MD simulation, the calculation of the relative free energies associated with different conformational states of a (bio)molecular system (and, possibly, of corresponding free-energy profiles or maps) is typically very sensitive to sampling errors. For example, the direct evaluation of the relative free energies corresponding to different structural motives of a solvated oligopeptide (e.g., various helix or sheet motives, folded and unfolded states) using plain MD simulations is only possible nowadays for short peptides,^{16–18} requires a sizable amount of computing time, and leads to results still affected by large uncertainties. These difficulties result directly from the long time scale associated with the corresponding interconversion processes, for example, for α -peptides in water based on refs 17, 19, and 20, α -helix formation²¹ (~ 200 ns), loop closing^{19,22} (~ 0.05 – 1 μ s), β -hairpin folding²³ (~ 1 – 10 μ s), and mini-protein folding²⁴ (~ 1 – 10 μ s).

In the numerous cases where a direct-counting approach^{16,25,26} based on plain MD simulations is not applicable, one may resort to umbrella sampling^{27,28} (US). The US approach relies on the use of a time-independent biasing potential in a relevant conformational subspace (US subspace) that forces the sampling of the conformational states under consideration with a sufficient number of interconversion transitions. In this case, the free-energy differences can be calculated from the ratio of the reweighted numbers of conformations assigned to the different states, the reweighting acting as a correction for the effect of the biasing. This approach is easily extended to the evaluation of free-energy profiles or maps. In practice, however, the direct design of a biasing potential

fulfilling the above conditions for significantly differing conformational states is difficult. There exist three basic approaches to overcome this problem.

In multiple-windows schemes,^{29–32} a set of MD simulations is performed using different local biasing potentials (typically harmonic) that restrict the sampling to specific regions of the US subspace. The data from the different (overlapping) biased ensembles is then assembled, either manually or automatically (e.g., using the weighted-histogram^{33–36} or umbrella-integration^{37,38} methods) to construct the complete free-energy profile or map. In adaptive (iterative) schemes,^{35,39–48} successive equilibrium MD simulations are used to progressively optimize a nonlocal biasing potential (typically nonharmonic) with the goal of achieving a nearly complete sampling of the US subspace. This can be done by examining the conformational probability at each step and adjusting the biasing potential for the next step so as to remove undersampled regions. The data from the different steps is then combined to construct the final free-energy profile or map. Finally, in memory-based schemes,^{15,49–52} the strategy is similar to that of the adaptive approach, but a single (relatively short) nonequilibrium (memory-building) MD simulation is used as a preoptimization tool for the nonlocal biasing potential, followed by a single (comparatively longer) equilibrium US simulation for production.

The last approach relies on methods previously developed to address the conformational searching problem,^{14,15} that is, the problem of scanning a potential energy hypersurface for low-energy configurations over the widest possible volume, without any requirement on the probability distribution of these configurations. Memory-based methods are probably the most efficient types of MD-based searching methods available nowadays. They rely on the progressive build-up of a time-dependent penalty potential, which prevents the continuous revisiting of previously discovered configurations. Many closely related variants of this approach can be found in the literature, including (chronologically) the deflation,⁵³ tunneling,⁵⁴ tabu search,⁵⁵ local elevation,⁵⁶ conformational flooding,⁵⁷ Engkvist–Karlström,⁵⁸ adaptive reaction coordinate force,⁴⁹ adaptive biasing force,⁵² metadynamics,^{50,51} and filling potential⁵⁹ methods. The first practically useful implementation of a memory-based searching scheme in the context of (bio)molecular systems with explicit solvation is probably the local elevation (LE) method of Huber, Torda, and van Gunsteren,⁵⁶ as implemented in the GROMOS96 program.^{60,61} In this method, the searching enhancement is applied along a subset of degrees of freedom of the system (LE subspace), typically a limited set of conformationally relevant dihedral angles, by means of a penalty potential defined as a weighted sum of local (grid-based, short-ranged) repulsive Gaussian functions, the corresponding weights being made proportional to the number of previous visits to the specific conformation (grid cell). Because memory-based searching methods (such as the LE method) have a time-dependent Hamiltonian, they sample in principle no well-defined configurational probability distribution, that is, the resulting trajectories cannot be used for the evaluation of thermodynamic properties, including free energies, via statistical mechanics. However, in view of their

very high searching power, there has been a long-standing interest in using their basic principle to design efficient conformational sampling methods, that is, leading to trajectories suited for the evaluation of thermodynamic properties after reweighting. This can be done by observing that at the end of a memory-based search, the penalty potential has essentially “flattened” the free-energy surface in the considered subspace up to a certain threshold value above the lowest minimum discovered.⁵⁸ As a result, this final penalty potential represents an optimal biasing potential for a subsequent US simulation.

Such a combination is at the heart of the local elevation umbrella sampling^{15,62} (LEUS) approach. This scheme consists of two steps: (i) a LE build-up (searching) phase, that is used to construct an optimized memory-based biasing potential within a LE subspace of N_{LE} conformationally relevant degrees of freedom; (ii) an US sampling phase, where the (frozen) memory-based potential is used to generate a biased ensemble with extensive coverage of the US subspace defined by the same $N_{US} = N_{LE}$ degrees of freedom. A successful build-up phase will produce a biasing potential resulting in a nearly homogeneous coverage of the relevant subspace (up to a given free-energy level) during the subsequent sampling phase. Thermodynamic information appropriate for the physical (unbiased) ensemble can then be recovered from the simulated data by means of a simple reweighting procedure.^{15,27,28} Note that the LEUS approach bears some analogies with other memory-based sampling approaches^{14,15} such as the adaptive umbrella sampling,^{40,41,63} adaptive biasing force,^{52,64,68} adaptive reaction coordinate force,^{49,66–68} and metadynamics^{50,51} methods, but within a two-step implementation where the US sampling phase is used to correct for the inaccuracy of the (nonequilibrium) LE build-up phase, leading to a number of advantages in terms of efficiency, accuracy, and robustness.¹⁵ A similar two-step approach can also be found in related schemes developed by Babin et al.,^{44,48} Ensing et al.,⁶⁹ and Li et al.⁷⁰

The LEUS approach was previously applied to the calculation of the relative free energies of β -D-glucopyranose ring conformers in water,¹⁵ to the calculation of Ramachandran free-energy maps for the 11 glucose-based disaccharides in water,⁶² and, more recently, to the parametrization of a new GROMOS carbohydrate force field.⁷¹ This scheme was found to dramatically enhance the sampling power of MD simulations and to permit the calculation of accurate free-energy differences between relevant conformational states (as well as free-energy profiles or maps) for LEUS subspaces of low dimensionalities ($N_{LE} = N_{US} = 1-4$). This approach is efficient, nearly all the computational effort being invested in the actual sampling phase rather than in searching and equilibration, and robust, the method being only weakly sensitive to the details of the build-up protocol.^{15,62}

The LEUS approach represents a powerful sampling-enhancement tool in cases where the relevant conformational subspace is of low dimensionality. However, it becomes inapplicable as such for high-dimensional problems. Consider, for example, a decapeptide with 20 relevant degrees of freedom (successive ϕ and ψ backbone dihedral angles). If each degree of freedom is discretized into 32 bins (as in

refs 15 and 62) and if the biasing potential is expected to map out about 50% of the relevant conformational subspace, the number of local functions required is approximately 10^{30} , which is clearly intractable in terms of both memory and build-up duration requirements.

A tentative solution to this problem relies on the optimization of fragment-based biasing potentials of low dimensionalities (e.g., in the peptide case, two-dimensional potentials optimized for the ϕ and ψ dihedral angles corresponding to a specific type of residue pair, with a build-up involving the corresponding dipeptide, that is, $N_{LE} = 2$) and their simultaneous application to each of the corresponding fragments in a molecule (e.g., 10 successive ϕ and ψ pairs along the peptide backbone, that is, $N_{US} = 20$), following a similar principle as that developed in refs 72 and 73. The resulting combined biasing potential would eliminate the influence of the local conformational preferences of the successive fragments, without affecting nonlocal (longer-ranged) influences.

Such an extension of the LEUS approach to solvated polymers with more than a few relevant degrees of freedom is the goal of the present work. The resulting method will be termed fragment-based LEUS (FB-LEUS). More specifically, the FB-LEUS scheme involves an US sampling phase that relies on a superposition of low-dimensionality biasing potentials optimized using LEUS at the fragment level, that is, a situation with $N_{US} = N_F N_{LE}$, where N_F is the number of fragments in the molecule considered. This new combination appears to be very versatile, because optimized biasing potentials can in principle be precalculated for fragments (e.g., in the peptide case, ϕ and ψ backbone dihedral angles for all possible dipeptides), stored in a database, and later applied to larger molecules (e.g., on the corresponding dipeptide units of an oligopeptide with specified sequence), so as to enhance the sampling with a limited additional computational and memory cost.

The feasibility of the FB-LEUS approach is investigated here in the context of polyalanine (poly-Ala) and polyvaline (poly-Val) oligopeptides in water. Two-dimensional biasing potentials are preoptimized at the mono-peptide level, distinguishing between alanine and valine, as well as between N-blocked, C-blocked, and N&C-blocked mono-peptides, assumed representative for the C-terminal, N-terminal, and intermediate residues of the oligopeptides, respectively. These potentials are then applied to all dihedral-angle pairs within (unblocked) oligopeptides of 4, 6, 8, or 10 residues. Two types of fragment-based biasing potentials are distinguished: (i) the basin-filling (BF) potentials act so as to “fill” free-energy basins up to a prescribed free-energy level above the global minimum; (ii) the valley-digging (VD) potentials act so as to “dig” valleys between the (four) relevant free-energy minima of the two-dimensional maps, preserving barriers (relative to linearly interpolated free-energy changes) of a prescribed magnitude. Two important differences between the BF and VD biasing potentials at the fragment level are that (i) the VD potential opens up a much smaller volume of irrelevant conformational space (i.e., conformations associated with low Boltzmann weights in the physical ensemble) and (ii) the VD potential preserves the relative

free energies of the (four) minima, while the BF potential equalizes their values. However, both potentials will promote an increased number of transitions between these minima. The results of the simulations involving different FB-LEUS biasing schemes are analyzed in terms of searching efficiency (volume of conformational space visited in a given amount of simulation time), statistical efficiency (representativeness of the biased ensemble in terms of the conformational distribution appropriate for the physical ensemble), and reliability of the predicted low free energy conformations and associated free energies (convergence with time, number of interconversion transitions, and mutual overlap across schemes).

2. Computational Details

2.1. Simulation Procedure. All MD simulations were carried out using a modified version of the GROMOS05 program,⁷⁴ together with the GROMOS 53A6 force field⁷⁵ and the simple-point charges (SPC) water model.⁷⁶ They were performed under periodic boundary conditions based on cubic computational boxes and in the isothermal–isobaric (NPT) ensemble at 300 K and 1 bar. Newton's equations of motion were integrated using the leapfrog algorithm^{77,78} with a 2 fs time step. The SHAKE procedure⁷⁹ was applied to constrain all bond lengths as well as the full rigidity of the solvent molecules with a relative geometric tolerance of 10^{-4} . The temperature was maintained by weakly coupling the solute and solvent degrees of freedom (jointly) to a temperature bath⁸⁰ at 300 K with a relaxation time of 0.1 ps. The pressure was maintained by weakly coupling the atomic coordinates and box dimensions (isotropic coordinate scaling, group-based virial) to a pressure bath⁸⁰ at 1 bar with a relaxation time of 0.5 ps and an isothermal compressibility of $0.4575 \times 10^{-3} \text{ (kJ mol}^{-1} \text{ nm}^{-3})^{-1}$ as appropriate for water.⁶⁰ The center of mass motion was removed every time step. The nonbonded interactions were handled using a twin-range cutoff scheme,^{2,81} based on short- and long-range cutoff distances of 0.8 and 1.4 nm, respectively, and an update frequency of 5 time steps for the short-range pairlist and intermediate-range interactions. The mean effect of the omitted electrostatic interactions beyond the long-range cutoff distance was approximately reintroduced using a reaction-field correction,⁸² with a relative dielectric permittivity of 61 as appropriate for the SPC water model.³¹ Values of the successive backbone ϕ and ψ dihedral angles, as well as of the biasing potential, were written to file every time step for subsequent analysis.

The simulated systems (computational box) consisted of one solute molecule and $N_{\text{H}_2\text{O}}$ water molecules. The solute molecules considered (Figure 1) were (partially) blocked alanine or valine mono-peptides (Ala_1^{N} , Ala_1^{C} , Ala_1^{NC} , Val_1^{N} , Val_1^{C} or Val_1^{NC} , along with $N_{\text{H}_2\text{O}} = 1300$) and unblocked alanine or valine oligopeptides (Ala_n or Val_n with $n = 4, 6, 8, \text{ or } 10$, along with $N_{\text{H}_2\text{O}} = 2100, 3500, 5500, \text{ and } 8000$, respectively). Note that the term “dipeptide” is sometimes used rather than “monopeptide” for the compounds Ala_1 and Val_1 , the latter term appearing more consistent to the authors. Unblocked termini were modeled using parameters appropri-

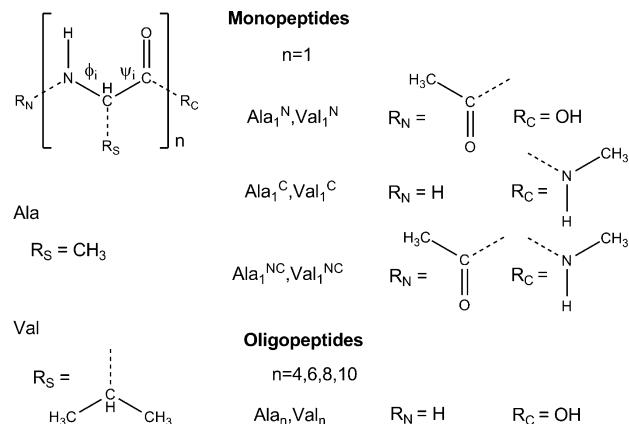


Figure 1. Solute molecules considered in the present study. These include (partially) blocked alanine or valine mono-peptides (Ala_1^{N} , Ala_1^{C} , Ala_1^{NC} , Val_1^{N} , Val_1^{C} or Val_1^{NC}) and unblocked alanine or valine oligopeptides (Ala_n or Val_n with $n = 4, 6, 8, \text{ or } 10$). Unblocked termini correspond to unprotonated amine (N-terminus) and protonated carboxylic acid (C-terminus) groups. Blocked termini correspond to acetylated (N-terminus) and methylamidated (C-terminus) groups.

ate for the uncharged forms of the corresponding free groups (unprotonated amine and protonated carboxylic acid). Blocked termini were modeled using parameters appropriate for the acetylated (N-terminus) and methylamidated (C-terminus) forms of these groups. Although the termini of unblocked peptides are expected to be ionized in aqueous solution at neutral pH, the choice was made here to consider uncharged termini for the oligopeptides Ala_n and Val_n ($n = 4, 6, 8, \text{ or } 10$), so as to avoid conformational ensembles dominated by (nonlocal) interactions between terminal charges.^{11,83} In addition to providing a simpler context for the testing of a fragment-based sampling-enhancement approach, this situation is also more representative of an experimental setup involving a blocked peptide, an excess of counterions or a peptide segment within a protein. The simulations were initiated from fully extended peptide structures, subsequently relaxed by 500 ps MD equilibration.

The oligopeptide simulations (Ala_n or Val_n with $n = 4, 6, 8, \text{ or } 10$) were carried out using the proposed FB-LEUS approach. Corresponding unbiased simulations were also undertaken for comparison. The fragment-based biasing potentials were constructed at the monopeptide level (Ala_1^{N} , Ala_1^{C} , Ala_1^{NC} , Val_1^{N} , Val_1^{C} , and Val_1^{NC}) following different approaches (section 2.2). The design of these potentials relied on Ramachandran free-energy maps for the monopeptides, evaluated using the standard LEUS method.¹⁵ Note that truncated-polynomial local functions (of widths equal to the grid spacing) were used in the present work rather than the previously employed Gaussian functions,^{15,56,62} as described in Appendix A.

2.2. FB-LEUS Schemes. If the essentially stiff peptide bonds are omitted, the backbone conformation of a mono- or oligopeptide, Ala_n or Val_n ($n = 1, 4, 6, 8, \text{ or } 10$, irrespective of its terminal blocking), is defined by n pairs of dihedral angles ϕ_i and ψ_i with $i = 1, \dots, n$, numbered starting from the N-terminus and simply noted ϕ and ψ for $n = 1$ (Figure 1).

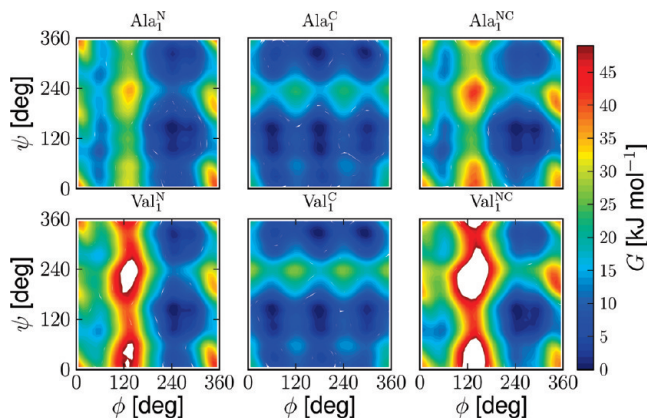


Figure 2. Ramachandran free-energy maps, $G(\phi, \psi)$, for the mono-peptides (Ala_1^{N} , Ala_1^{C} , Ala_1^{NC} , Val_1^{N} , Val_1^{C} , or Val_1^{NC} , Figure 1), evaluated using LEUS simulations ($t_{\text{LE}} = 15$ ns for alanine or 20 ns for valine, $t_{\text{US}} = 50$ ns). All maps are anchored by the condition $G(\phi, \psi) = 0$ at the global minimum. Regions displayed in white are associated with very high relative free energies (>50 kJ mol^{-1}) due to steric clashes. Note that the maps are drawn considering $[0^\circ, 360^\circ]$ dihedral-angle ranges rather than the more usual $[-180^\circ, 180^\circ]$ ranges.

Simulations of the six (partially) blocked alanine or valine mono-peptides were used to preoptimize a library of fragment-based two-dimensional biasing potentials. To this purpose, Ramachandran free-energy maps, $G(\phi, \psi)$, were first evaluated for the six compounds using the LEUS method,¹⁵ with a dimensionality $N_{\text{LE}} = N_{\text{US}} = 2$ (ϕ and ψ), a number of grid points per dimension $N_{\text{G}} = 32$ (angular spacing 11.25°), a force-constant increment per visit $k_{\text{LE}} = 2 \times 10^{-3}$ kJ mol^{-1} , a build-up duration $t_{\text{LE}} = 15$ ns (alanine) or 20 ns (valine), and a sampling duration $t_{\text{US}} = 50$ ns (Appendix A). The resulting maps, obtained from the sampling phases of these simulations after application of the proper reweighting¹⁵ are displayed in Figure 2. These maps were calculated with the same grid spacing as used in the LEUS scheme and anchored by the condition $G(\phi, \psi) = 0$ at the global minimum. The maps for N-blocked and N&C-blocked mono-peptides present four free-energy basins centered at ϕ values of about 60° and 260° , along with ψ values of about 120° and 300° , the main (deepest and widest) basin corresponding to $(\phi, \psi) = (260^\circ, 120^\circ)$. For these mono-peptides, the rotation around ψ is associated with low barriers (similar for alanine and valine), while the rotation around ϕ is associated with a higher barrier at $\phi \approx 120^\circ$ (significantly higher for valine compared with alanine). In contrast, the maps for C-blocked mono-peptides present six free-energy basins centered at ϕ values of about 60° , 180° , and 300° , along with ψ values of about 120° and 300° , all of comparable depths and widths. For these mono-peptides, the rotations around both ϕ and ψ are associated with low barriers (similar for alanine and valine). The fragment-based biasing potentials were derived from these maps according to two different procedures.

The basin-filling (BF) biasing potentials are defined as

$$\mathcal{U}_{\text{bias}}^{\text{BF}}(\phi, \psi; h) = \begin{cases} h - G(\phi, \psi) & \text{if } G(\phi, \psi) \leq h \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

where h denotes a free-energy level (relative to the global minimum of the map) up to which the biased map $G + \mathcal{U}_{\text{bias}}^{\text{BF}}$ is “flattened”. In practice, $\mathcal{U}_{\text{bias}}^{\text{BF}}$ is constructed from the calculated free-energy surface by using a grid-based version of eq 1, as detailed in Appendix B. The BF approach is illustrated schematically in Figure 3a. Three values of h were considered, $h = 10, 20$, or 30 kJ mol^{-1} , resulting in BF biasing potentials labeled F_{10} , F_{20} , and F_{30} , respectively. These three types of biasing potentials were evaluated separately for Ala_1^{N} , Ala_1^{C} , Ala_1^{NC} , Val_1^{N} , Val_1^{C} , and Val_1^{NC} . Those for Ala_1^{NC} and Val_1^{NC} are illustrated in Figure 3b in the form of biased maps $G + \mathcal{U}_{\text{bias}}^{\text{BF}}$.

The valley-digging (VD) biasing potentials are defined as

$$\mathcal{U}_{\text{bias}}^{\text{VD}}(\phi, \psi; h) = \begin{cases} h - \Delta G_k(\phi, \psi) & \text{if } (\phi, \psi) \rightarrow k \text{ and } \Delta G_k(\phi, \psi) \geq h \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

Here, the biasing potential is defined in terms of eight line segments k , each extending along either ϕ or ψ , and connecting the points (ϕ, ψ) of the set $\{(61.875^\circ, 118.125^\circ), (275.625^\circ, 118.125^\circ), (61.875^\circ, 298.125^\circ), (275.625^\circ, 298.125^\circ)\}$, four of them directly and four of them across a period. These points belong to the LEUS grid, so that the connecting segments encompass a series of grid points. For a segment k extending from $\phi_k^{(b)}$ to $\phi_k^{(e)}$ along ϕ at ψ_k , the quantity $\Delta G_k(\phi, \psi)$ in eq 2 is defined as

$$\Delta G_k(\phi, \psi) = G(\phi, \psi) - \{[1 - \lambda_k(\phi)]G(\phi_k^{(b)}, \psi_k) + \lambda_k(\phi)G(\phi_k^{(e)}, \psi_k)\} \quad (3)$$

where

$$\lambda_k(\phi) = \frac{\phi - \phi_k^{(b)}}{\phi_k^{(e)} - \phi_k^{(b)}} \quad (4)$$

and the notation $(\phi, \psi) \rightarrow k$ denotes a point belonging to segment k , that is, satisfying

$$2|\psi - \psi_k| \leq \sigma \quad \text{and} \quad \phi_k^{(b)} + |\psi - \psi_k| < \phi < \phi_k^{(e)} - |\psi - \psi_k| \quad (5)$$

Similar definitions apply to a segment k extending from $\psi_k^{(b)}$ to $\psi_k^{(e)}$ along ψ at ϕ_k . When a point belongs to segment k , λ_k measures its fractional longitudinal distance from the beginning point of the segment relative to the full segment length, and $\Delta G_k(\phi, \psi)$ is the difference between the free energy at this point and a free energy value linearly interpolated from those at the beginning and end points of the segment. Setting h to zero will lead to a biased map $G + \mathcal{U}_{\text{bias}}^{\text{VD}}$ presenting a connection of the four above points via eight line segments associated with linearly varying free energies (valleys). The actual value of h denotes a free-energy level (relative to this linear free-energy variation) down to which the valleys are “dug”. Note that the second condition in eq 5 is formulated in such a way that a given (ϕ, ψ) point can belong to at most one segment, even close to the connecting points, that is, that the eight segments are nonoverlapping. In practice, $\mathcal{U}_{\text{bias}}^{\text{VD}}$ is constructed from the calculated free-energy surface by using grid-based versions of eqs 2–5, as detailed in Appendix B. In

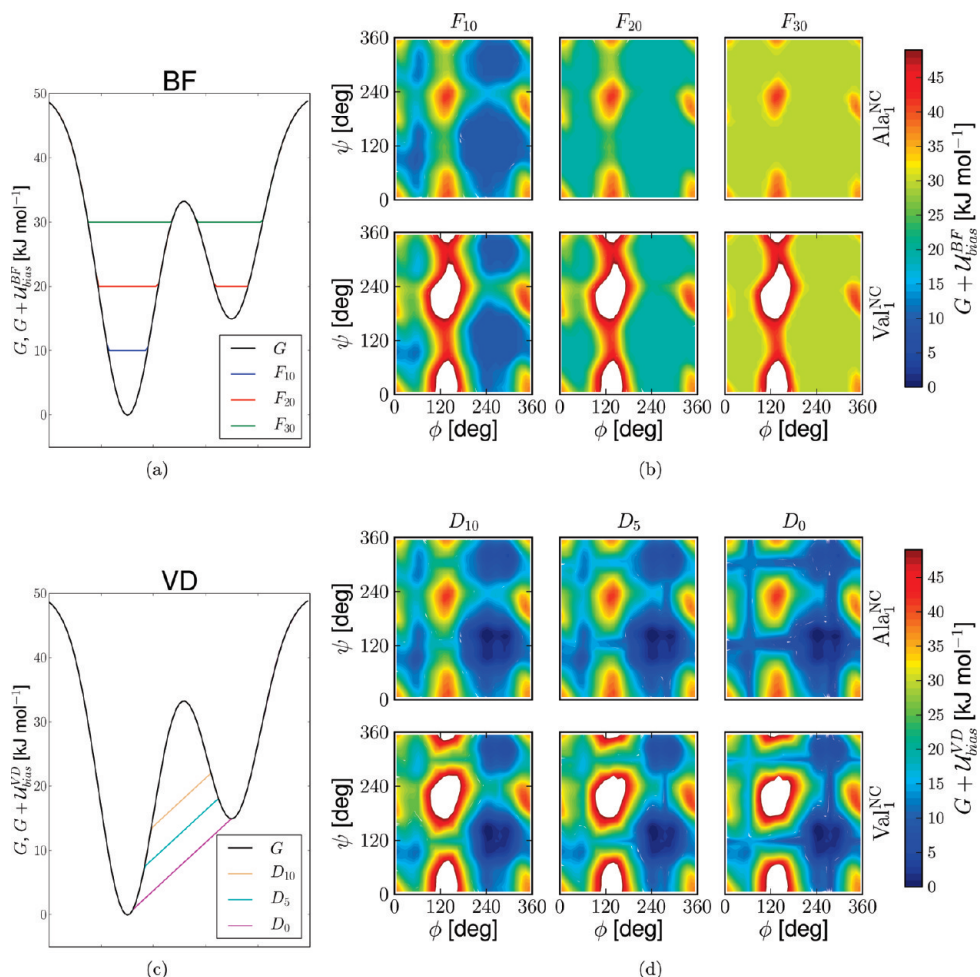


Figure 3. Schematic illustrations of the BF and VD procedures for generating fragment-based biasing potentials and corresponding biased maps for the mono-peptides Ala_1^{NC} and Val_1^{NC} (Figure 1) associated with different parameters h of the BF and VD potentials. The biased maps are defined by $G + \mathcal{U}_{\text{bias}}$, where $G(\phi, \psi)$ is the free energy of the physical system (Figure 2) and $\mathcal{U}_{\text{bias}}(\phi, \psi; h)$ is the biasing potential (eqs 1 or 2). The biasing potentials F_{10} , F_{20} , and F_{30} correspond to the BF procedure with $h = 10$, 20, or 30 kJ mol^{-1} , respectively. The biasing potentials D_{10} , D_5 , and D_0 correspond to the VD procedure with $h = 10$, 5, or 0 kJ mol^{-1} , respectively. Note that the maps are drawn considering $[0^\circ, 360^\circ]$ dihedral-angle ranges rather than the more usual $[-180^\circ, 180^\circ]$ ranges.

addition, to avoid spurious transverse oscillatory motions within the segments, the grid-based functions corresponding to all lines parallel to the eight above segments are adjusted so as to level off the component of the biasing force orthogonal to the lines at a maximal magnitude of $1 \text{ kJ mol}^{-1} \text{ deg}^{-1}$. The VD approach is illustrated schematically in Figure 3c. Three values of h were considered, $h = 10$, 5, or 0 kJ mol^{-1} , resulting in VD biasing potentials labeled D_{10} , D_5 , and D_0 , respectively. These three types of biasing potentials were evaluated separately for Ala_1^{N} , Ala_1^{C} , Ala_1^{NC} , Val_1^{N} , Val_1^{C} , and Val_1^{NC} . Those for Ala_1^{NC} and Val_1^{NC} are illustrated in Figure 3d in the form of biased maps $G + \mathcal{U}_{\text{bias}}^{\text{VD}}$.

For comparison purposes, unbiased simulations were also undertaken. The corresponding “zero” biasing potential will be noted U (unbiased, an equivalent notation would be F_0). In this case, the biased maps analogous to those reported in Figure 3 for the BF and VD potentials evaluated for Ala_1^{NC} and Val_1^{NC} are simply the unbiased maps of Figure 2 (right panels).

After the definition of the different fragment-based biasing potentials, these were applied in simulations of the longer

(unblocked) oligopeptides Ala_n and Val_n ($n = 4, 6, 8$, or 10) in the following way. The potentials derived for Ala_1^{C} and Val_1^{C} were applied to (ϕ_1, ψ_1) in Ala_n and Val_n , respectively, the potentials derived for Ala_1^{N} and Val_1^{N} to (ϕ_n, ψ_n) in Ala_n and Val_n , respectively, and the potentials derived for Ala_1^{NC} and Val_1^{NC} to all other dihedral angles (ϕ_i, ψ_i) with $i = 2, \dots, n - 1$ in Ala_n and Val_n , respectively. Note that if, for simplicity, the same terminal blocking groups have been used here for Ala_1 and Val_1 , the use of groups involving an isopropyl (rather than a methyl) termination would possibly be more appropriate for Val_1 , being more representative for $\text{Val}-\text{Val}$ interactions within a peptide. An alternative approach would involve unblocked Ala_2 and Val_2 fragments, with ψ_1 and ϕ_2 taken as representative for (ψ_i, ϕ_{i+1}) with $i = 1, \dots, n - 1$ within Ala_n and Val_n , respectively (leaving ϕ_1 and ψ_n unbiased).

The different FB-LEUS simulations were carried out for a sampling duration $t_{\text{US}} = 50 \text{ ns}$ (100 ns for the unbiased simulations). This resulted in a set of 56 simulations, depending on the type of oligopeptide considered (Ala or

Val), number of residues ($n = 4, 6, 8,$ or 10), and type of biasing potential ($F_{10}, F_{20}, F_{30}, D_{10}, D_5, D_0,$ or U).

2.3. Analysis Procedure. The overall efficiency achieved by a specific sampling-enhancement scheme is a combination of two factors: (i) the searching efficiency, that is, the ability of the scheme to search for low free energy regions across a wide extent of conformational space, thereby escaping local free-energy minima and overcoming free-energy barriers; (ii) the statistical efficiency, that is, the fraction of the sampled configurations actually relevant in terms of the conformational (Boltzmann) distribution characteristic of the physical system.

The searching efficiency was assessed for the Ala_{*n*} and Val_{*n*} oligopeptides by monitoring the cumulative number $N_n(t)$ of unique peptide backbone conformations discovered up to time t during the sampling phases of the simulations involving different biasing potentials. In the present study, a unique backbone conformation is defined by an integer corresponding to a string of $2n$ bits (arranged from the highest-weight to the lowest-weight bit). Given a numbering of the backbone dihedral angles from the N-terminus to the C-terminus of the peptide, bit $2i - 1$ ($i = 1, \dots, n$) is set to zero if the backbone dihedral angle ϕ_i is in the range $[130^\circ; 360^\circ]$ and to one otherwise, while bit $2i$ ($i = 1, \dots, n$) is set to zero if the backbone dihedral angle ψ_i is in the range $[0^\circ; 230^\circ]$ and to one otherwise. These intervals were selected by consideration of the free-energy maps for the mono-peptides as approximately defining four distinct free-energy basins (Figure 2). At the mono-peptide level ($n = 1$), the bit strings 00 and 01 correspond to the sheet and helical regions of the Ramachandran map, respectively, while 10 and 11 correspond to typically less populated regions (in proteins). Based on this assignment, there are in total $N_n^{\max} = 4^n$ unique backbone conformations for Ala_{*n*} or Val_{*n*}, defining the corresponding exhaustive-search upper bound for $N_n(t)$. The different $N_n(t)$ curves were fitted to stretched-exponential functions of the form

$$N_n(t) = N_n^{\max} \{1 - \exp[-(\tau_n^{-1}t)^{\alpha_n}]\} \quad (6)$$

The resulting parameters α_n and τ_n are further referred to as the stretching exponent and the characteristic searching time, respectively, of the simulation involving a specific biasing potential. The stretching exponent α_n is expected to be one for trajectories representing an entirely random configuration generation process. Negative deviations ($0 < \alpha_n < 1$) account for two effects: (i) the presence of time correlations in real trajectories, which represent a diffusion process in configuration space; (ii) the presence of a probability bias in real trajectories, which generate configurations according to a (possibly biased) Boltzmann distribution in configuration space. Although the two effects are not clearly separable in practice in terms of their influence on the searching rate, they both induce a tendency of the system to revisit previously discovered configurations, leading to a slower evolution of $N_n(t)$ toward N_n^{\max} . Irrespective of the value of α_n , the characteristic searching time τ_n represents the time required for the trajectory to cover a fraction $1 - e^{-1}$ (63%) of the entire configuration space accessible to the system.

This value can be reexpressed in terms of an effective visiting time $\tilde{\tau}_n$, defined as

$$\tilde{\tau}_n = \frac{\tau_n}{(1 - e^{-1})N_n^{\max}} \quad (7)$$

Under the assumption of an entirely random configuration generation process (i.e., when $\alpha_n = 1$), $\tilde{\tau}_n$ can be interpreted as the average time separating the visit of two conformations (including newly discovered and revisited ones) or, equivalently, as the average time the trajectory spends in a given conformation before leaving it.

The statistical efficiency was assessed for the Ala_{*n*} and Val_{*n*} oligopeptides by calculating, for the simulations involving different biasing potentials, the quantity F_n defined as⁶²

$$F_n = N_f^{-1} \exp[-\sum_k p_k \ln p_k] \quad (8)$$

where N_f is the number of considered trajectory frames (25×10^6 for the 50 ns sampling phases of the present simulations with configurations stored every 2 fs) and p_k is the statistical (unbiasing) weight associated with frame k , defined as

$$p_k = \left[\sum_{l=1}^{N_f} \exp[\beta \mathcal{U}_{\text{bias},l}] \right]^{-1} \exp[\beta \mathcal{U}_{\text{bias},k}] \quad (9)$$

where $\beta = (k_B T)^{-1}$, k_B being Boltzmann's constant and T the absolute temperature (300 K), and $\mathcal{U}_{\text{bias},k}$ is the value of the biasing potential associated with trajectory frame k . The limiting case of an unbiased simulation corresponds to $p_k = N_f^{-1}$ for all frames, leading to $F_n = 1$ (maximum statistical efficiency). The limiting case of a biased simulation where a single frame k entirely dominates the reweighted probability distribution corresponds to $p_l = 0$ for $l \neq k$ along with $p_k = 1$, leading to $F_n = N_f^{-1}$ (minimum statistical efficiency, very close to zero). The quantity $F_n N_f$ can thus be viewed as an effective number of frames of the biased trajectory contributing to the statistics in terms of the properties of the unbiased ensemble. Although one usually has $F_n \ll 1$ in a biased simulation, the sampling efficiency may still be greatly enhanced in practice when this effective number of frames spans a much wider (i.e., more representative) volume of the configuration space accessible to the unbiased system, that is, when the searching efficiency has been increased. The significance of the statistical efficiency factor F_n is discussed in more detail in Appendix C. Note that a number of other measures for the statistical efficiency have been proposed previously.^{25,84-87}

Considering the above discussion, the opposing effects of an enhancement of the searching efficiency and a deterioration of the statistical efficiency upon applying a biasing potential can be characterized by means of a combined sampling efficiency parameter $C_n(t_0)$ associated to a given time scale t_0 , defined as

$$C_n(t_0) = \tilde{\tau}_n^{-1} \frac{1 - \exp[-(\tau_n^{-1}t_0)^{\alpha_n}]}{1 - \exp[-(\tau_n^{-1}t_0)]} F_n \quad (10)$$

The first factor accounts for the rate at which successive conformations (including newly discovered and revisited ones) are produced (eq 7). The second factor accounts for the tendency to revisit known conformations on a time scale t_0 , compared to a purely random configuration generation process (eq 6). The third factor accounts for the statistical efficiency, that is, the relevance of the generated configurations in terms of the physical ensemble (eq 8). In practice, because a sampling enhancement or deterioration is always defined by reference to plain MD simulation for a given system, the combined sampling efficiency parameter will be reported as $\tilde{C}_n(t_0)$, defined as the quotient of $C_n(t_0)$ for a specific biasing scheme to the corresponding value for the unbiased simulation.

For practical applications, a more directly relevant point to be addressed concerns the reliability with which the low free energy conformations and associated free energies can be determined using a given sampling scheme. To this purpose, the lowest free energy conformations predicted by a given sampling scheme were also analyzed in terms of (i) convergence with time, (ii) number of interconversion transitions, and (iii) mutual overlap across schemes.

As a measure for the convergence of the calculated free energies over time, a convergence score $S_n(t)$ was defined as

$$S_n(t) = \left\{ \sum_{m=1}^{N_m^{\max}} [\exp[-\beta G_m(t)] - \exp[-\beta G_m(t_s)]]^2 \right\}^{1/2} \quad (11)$$

where t_s is a reference total simulation time (set to 50 ns for all simulations), and $G_m(t)$ is the free energy predicted for backbone conformation m based on the simulation time t , defined as

$$G_m(t) = -\beta^{-1} \ln \left\{ \left[\sum_{l=1}^{N_f(t)} \exp[\beta \mathcal{U}_{\text{bias},l}] \right]^{-1} \sum_{k=1, k \in \mathcal{S}_m}^{N_f(t)} \exp[\beta \mathcal{U}_{\text{bias},k}] \right\} \quad (12)$$

where $N_f(t)$ denotes the number of trajectory frames generated up to time t and \mathcal{S}_m the subset of these trajectory frames assigned to conformation m . Note that any conformation that has not been sampled at time t or at time t_s is characterized by an infinite free energy and leads to a corresponding exponential factor of zero in eq 11. The score $S_n(t)$ accounts for the root-sum-square error in the predicted conformational populations for a simulation of duration t compared with the full simulation of duration t_s . Because population differences will be largest for the conformations that are most populated at time t or at time t_s , this score gives more weight to unconverged free-energy estimates within the lowest free energy conformations. The convergence score will typically evolve from a value somewhat below one for $t \approx 0$ to exactly zero at $t = t_s$ and indicates how well the predicted lowest free energy conformations and associated free energies are converged at time t .

As a measure for the number of interconversion transitions between the lowest free energy conformations, an average number of (direct or indirect) transitions $T_n(G_c)$ was moni-

tored, considering the $L_n(G_c)$ conformations within a free-energy cutoff G_c of the lowest free energy conformation discovered. More precisely, after evaluation of the relative free energies of the different conformations based on the entire simulation (after reweighting), the set of $L_n(G_c)$ lowest free energy conformations was identified. The number of transitions occurring along the trajectory between any two distinct conformations of this set (possibly via other conformations not included in the set) was then calculated and divided by $L_n(G_c)$, leading to the average number of transitions per conformation $T_n(G_c)$. This number provides an indication concerning the expected accuracy with which the relative free energies of the low free energy conformations are estimated, because the relative populations of two states are only likely to be representative of an equilibrium situation if interconversions are frequent on the simulation time scale. The calculation of $T_n(G_c)$ was performed separately for $G_c = 5, 10, \text{ or } 15 \text{ kJ mol}^{-1}$ based on 50 ns simulations (first 50 ns for U).

Finally, as a measure for the (dis)agreement between the lowest free energy conformations generated using different schemes, overlap measures $O_n^{AB}(G_c)$ and $O_n^{BA}(G_c)$ were monitored for all pairs of simulations (A and B) involving different biasing potentials, considering the corresponding sets of $L_n^A(G_c)$ and $L_n^B(G_c)$ conformations within a free-energy cutoff G_c of the lowest free energy conformation discovered. The overlaps $O_n^{AB}(G_c)$ and $O_n^{BA}(G_c)$ are defined as the number of common configurations of the two sets, divided by $L_n^A(G_c)$ or $L_n^B(G_c)$, respectively. These overlap measures are comprised between 0 (no overlap) and 1 ($O_n^{AB}(G_c)$, set B encompasses all conformations of set A ; $O_n^{BA}(G_c)$, set A encompasses all conformations of set B ; both, full overlap). The calculation was performed separately for $G_c = 5, 10, \text{ or } 15 \text{ kJ mol}^{-1}$ based on 50 ns simulations (first 50 ns for U).

Finally, to further illustrate the extent of overlap between the different simulations, a set of five consensus lowest free energy conformations was identified, considering simultaneously the seven 50 ns simulations (first 50 ns for U) performed for each of the eight systems (Ala_n or Val_n with $n = 4, 6, 8, \text{ or } 10$). For a given system, each backbone conformation m was attributed a consensus rank R_m defined by

$$R_m = \sum_{i=1}^7 (N_i - R_{m,i}) \exp[-\beta G_{m,i}] \quad (13)$$

where N_i is the number of conformations visited in simulation i , $R_{m,i}$ is the free-energy rank of conformation m in this simulation, and $G_{m,i}$ the corresponding free energy (the latter two measured relative to the minimum free energy conformation predicted by this simulation). The five configurations with the lowest consensus rank were then identified, and their ranks $R_{m,i}$ and free energies $G_{m,i}$ in the seven individual simulations are reported.

3. Results

The curves representing the cumulative number $N_n(t)$ of unique peptide backbone conformations discovered up to

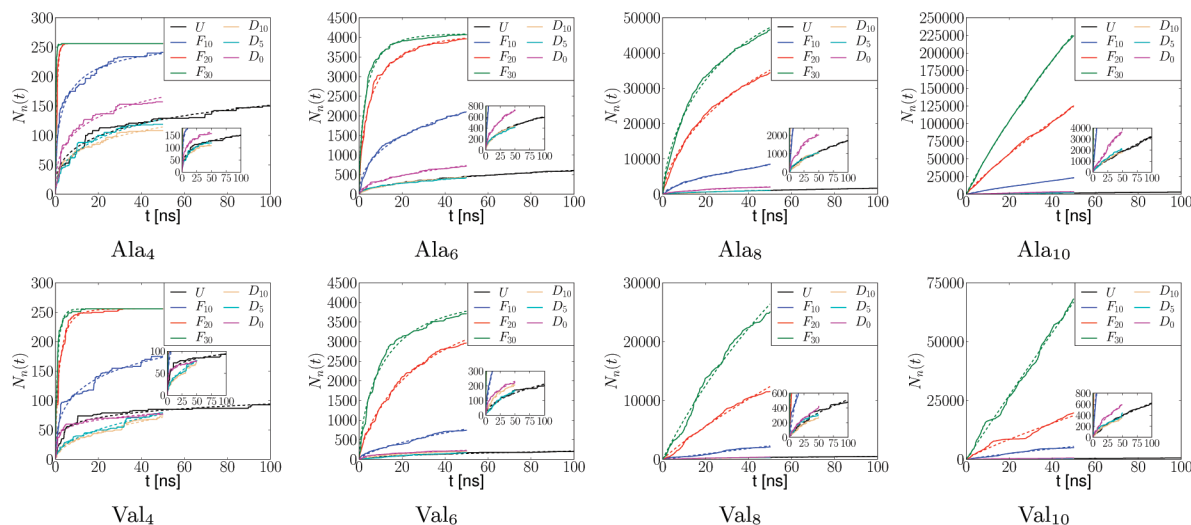


Figure 4. Number, $N_n(t)$, of unique peptide backbone conformations visited as a function of the sampling time t for Ala_n and Val_n oligopeptide simulations performed with different FB-LEUS biasing potentials. The biasing potentials considered are illustrated in Figure 3. The $N_n(t)$ curves are displayed using solid lines. Stretched-exponential fits (eqs 6 and 7) are displayed using dashed lines, the corresponding parameters α_n , τ_n and $\tilde{\tau}_n$ being reported in Table 1 and displayed graphically in Figure 5a,b ($\alpha_n, \tilde{\tau}_n$).

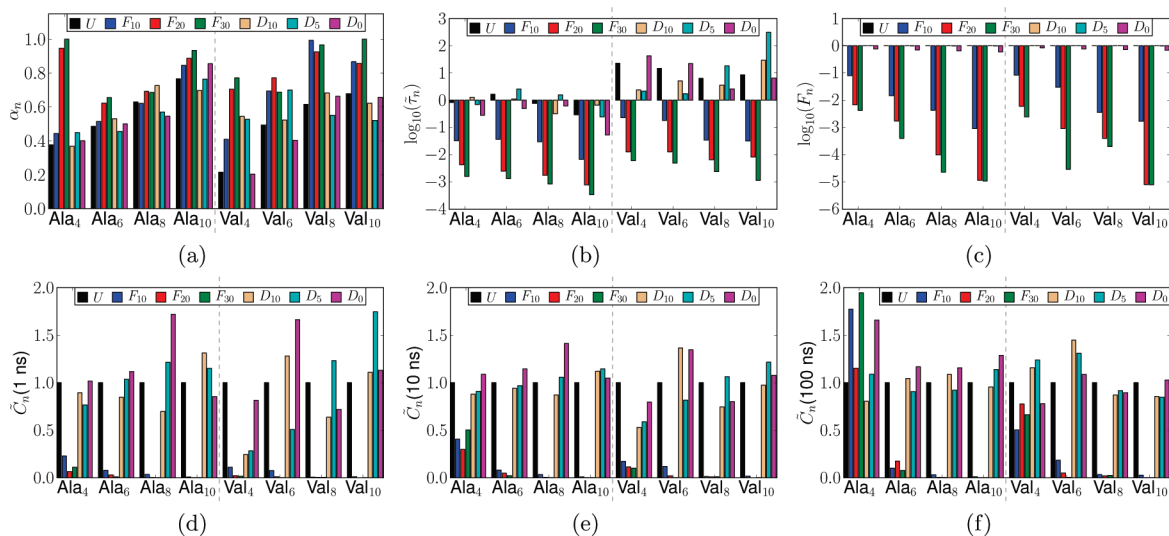


Figure 5. Stretched exponent α_n , effective visiting time $\tilde{\tau}_n$, statistical efficiency F_n , and combined sampling efficiency $\tilde{C}_n(t_0)$ for Ala_n and Val_n oligopeptide simulations performed with different FB-LEUS biasing potentials. The biasing potentials considered are illustrated in Figure 3. The parameters α_n , $\tilde{\tau}_n$, F_n , and $\tilde{C}_n(t_0)$ are defined by eqs 6–10, where $\tilde{C}_n(t_0)$ is the ratio of $C_n(t_0)$ for a given simulation to the corresponding value for the unbiased simulation U , and reference times $t_0 = 1, 10$, or 100 ns are considered. Note that the quantities $\tilde{\tau}_n$ (in units of ns) and F_n are displayed on a (decimal) logarithmic scale. Corresponding numerical values can be found in Table 1.

time t during the sampling phase of the different Ala_n and Val_n oligopeptide simulations are displayed in Figure 4. In terms of searching efficiency, the application of the FB-LEUS biasing potentials may lead to a spectacular enhancement compared with the unbiased MD simulation U .

For all oligopeptides, this enhancement is most pronounced when using the BF biasing potentials, systematically increasing along the series F_{10} , F_{20} , and F_{30} of increasingly “flattened” monopeptide free-energy surfaces (Figure 3a,b). For Ala_4 and Val_4 , the exhaustive-search upper bound $N_4^{\text{max}} = 256$ for $N_4(t)$ is reached within 2.8 (Ala_4) or 13.8 (Val_4) ns when the biasing potential F_{30} is used and within 4.4 (Ala_4) or 31.4 (Val_4) ns when the biasing potential F_{20} is used, while plain MD only visits 150 (Ala_4) or 93 (Val_4) conformations

within 100 ns. For Ala_6 , the corresponding upper bound $N_6^{\text{max}} = 4096$ for $N_6(t)$ is also reached close to 50 ns when using the biasing potential F_{30} , while plain MD only visits 598 conformations within 100 ns. Considering all systems, the numbers of conformations that have been visited within 50 ns is increased by factors of 1.9–12.7 (F_{10}), 2.0–65.6 (F_{20}), and 2.0–172.9 (F_{30}) compared to the unbiased simulation U , these factors systematically increasing with the oligopeptide length n and being tendentially larger for Val_n compared with Ala_n at a given n .

The searching enhancement is much less pronounced when the VD biasing potentials are used, tendentially increasing along the series D_{10} , D_5 , and D_0 of increasingly “dug” valleys on the monopeptide free-energy surfaces (Figure 3c,d).

Table 1. Stretching Exponent α_n , Characteristic Searching Time τ_n , Effective Visiting Time $\tilde{\tau}_n$, Statistical Efficiency F_n , and Combined Sampling Efficiency $\tilde{C}_n(t_b)$ for Ala_n and Val_n Oligopeptide Simulations Performed with Different FB-LEUS Biasing Potentials^a

bias	α_n	τ_n [ns]	$\tilde{\tau}_n$ [ns]	F_n	$\tilde{C}_n(1)$	$\tilde{C}_n(10)$	$\tilde{C}_n(100)$	bias	α_n	τ_n [ns]	$\tilde{\tau}_n$ [ns]	F_n	$\tilde{C}_n(1)$	$\tilde{C}_n(10)$	$\tilde{C}_n(100)$
Ala ₄															
U	0.38	1.3×10^2	8.3×10^{-1}	1.00×10^0	1.0000	1.0000	1.0000	U	0.22	3.7×10^3	2.3×10^1	1.00×10^0	1.0000	1.0000	1.0000
F ₁₀	0.44	5.2×10^0	3.2×10^{-2}	7.99×10^{-2}	0.2288	0.4057	1.7728	F ₁₀	0.41	3.7×10^1	2.3×10^{-1}	8.33×10^{-2}	0.1101	0.1726	0.5035
F ₂₀	0.95	7.0×10^{-1}	4.3×10^{-3}	6.73×10^{-3}	0.0655	0.2976	1.1525	F ₂₀	0.70	2.1×10^0	1.3×10^{-2}	6.01×10^{-3}	0.0218	0.1138	0.7763
F ₃₀	1.00	2.6×10^{-1}	1.6×10^{-3}	1.23×10^{-3}	0.1116	0.5024	1.9454	F ₃₀	0.77	9.9×10^{-1}	6.1×10^{-3}	2.44×10^{-3}	0.0158	0.1011	0.6634
D ₁₀	0.37	2.1×10^2	1.3×10^0	1.00×10^0	0.8943	0.8796	0.8040	D ₁₀	0.54	3.9×10^2	2.4×10^0	1.00×10^0	0.2444	0.5295	1.1568
D ₅	0.45	1.1×10^2	7.0×10^{-1}	9.89×10^{-1}	0.7656	0.9092	1.0881	D ₅	0.53	3.5×10^2	2.2×10^0	9.95×10^{-1}	0.2831	0.5878	1.2388
D ₀	0.40	4.6×10^1	2.8×10^{-1}	7.55×10^{-1}	1.0175	1.0882	1.6594	D ₀	0.20	6.8×10^3	4.2×10^1	8.38×10^{-1}	0.8149	0.7969	0.7791
Ala ₆															
U	0.49	4.3×10^3	1.7×10^0	1.00×10^0	1.0000	1.0000	1.0000	U	0.49	3.8×10^4	1.5×10^1	1.00×10^0	1.0000	1.0000	1.0000
F ₁₀	0.51	9.5×10^1	3.7×10^{-2}	1.47×10^{-2}	0.0792	0.0812	0.1014	F ₁₀	0.69	4.7×10^2	1.8×10^{-1}	2.99×10^{-2}	0.0761	0.1191	0.1847
F ₂₀	0.62	6.5×10^0	2.5×10^{-3}	1.72×10^{-3}	0.0292	0.0481	0.1757	F ₂₀	0.77	3.3×10^1	1.3×10^{-2}	9.15×10^{-4}	0.0110	0.0205	0.0505
F ₃₀	0.66	3.4×10^0	1.3×10^{-3}	3.98×10^{-4}	0.0097	0.0207	0.0769	F ₃₀	0.69	1.3×10^1	4.9×10^{-3}	2.92×10^{-5}	0.0099	0.0014	0.0043
D ₁₀	0.53	2.9×10^3	1.1×10^0	1.00×10^0	0.8470	0.9407	1.0443	D ₁₀	0.52	1.3×10^4	5.1×10^0	1.00×10^0	1.2808	1.3647	1.4474
D ₅	0.46	6.6×10^3	2.6×10^0	9.86×10^{-1}	1.0360	0.9678	0.9058	D ₅	0.70	4.5×10^3	1.7×10^0	9.93×10^{-1}	0.5069	0.8157	1.3100
D ₀	0.50	1.3×10^3	4.9×10^{-1}	6.96×10^{-1}	1.1170	1.1450	1.1678	D ₀	0.40	5.7×10^4	2.2×10^1	7.59×10^{-1}	1.6628	1.3459	1.0870
Ala ₈															
U	0.63	3.2×10^4	7.6×10^{-1}	1.00×10^0	1.0000	1.0000	1.0000	U	0.62	2.6×10^5	6.3×10^0	1.00×10^0	1.0000	1.0000	1.0000
F ₁₀	0.62	1.2×10^3	3.0×10^{-2}	4.30×10^{-3}	0.0352	0.0340	0.0322	F ₁₀	0.99	1.4×10^3	3.4×10^{-2}	3.59×10^{-3}	0.0057	0.0136	0.0326
F ₂₀	0.69	7.3×10^1	1.8×10^{-3}	9.83×10^{-5}	0.0034	0.0038	0.0049	F ₂₀	0.93	2.7×10^2	6.5×10^{-3}	3.98×10^{-4}	0.0048	0.0098	0.0200
F ₃₀	0.69	3.5×10^1	8.5×10^{-4}	2.29×10^{-5}	0.0013	0.0015	0.0023	F ₃₀	0.97	9.9×10^1	2.4×10^{-3}	1.99×10^{-4}	0.0051	0.0113	0.0256
D ₁₀	0.73	1.3×10^4	3.2×10^{-1}	1.00×10^0	0.6976	0.8709	1.0872	D ₁₀	0.68	1.5×10^5	3.6×10^0	1.00×10^0	0.6379	0.7448	0.8699
D ₅	0.57	6.5×10^4	1.6×10^0	9.84×10^{-1}	1.2146	1.0578	0.9217	D ₅	0.55	7.5×10^5	1.8×10^1	9.90×10^{-1}	1.2326	1.0629	0.9168
D ₀	0.55	2.6×10^4	6.2×10^{-1}	6.40×10^{-1}	1.7197	1.4129	1.1560	D ₀	0.66	1.1×10^5	2.6×10^0	7.20×10^{-1}	0.7190	0.8018	0.8937
Ala ₁₀															
U	0.77	1.9×10^5	2.9×10^{-1}	1.00×10^0	1.0000	1.0000	1.0000	U	0.68	5.7×10^6	8.6×10^0	1.00×10^0	1.0000	1.0000	1.0000
F ₁₀	0.85	4.5×10^3	6.7×10^{-3}	9.17×10^{-4}	0.0084	0.0101	0.0120	F ₁₀	0.87	2.1×10^4	3.2×10^{-2}	1.69×10^{-3}	0.0115	0.0177	0.0273
F ₂₀	0.89	5.2×10^2	7.8×10^{-4}	1.16×10^{-5}	0.0005	0.0007	0.0009	F ₂₀	0.86	5.5×10^3	8.2×10^{-3}	8.02×10^{-6}	0.0002	0.0003	0.0004
F ₃₀	0.93	2.3×10^2	3.4×10^{-4}	1.09×10^{-5}	0.0008	0.0011	0.0017	F ₃₀	1.00	7.6×10^2	1.1×10^{-3}	7.92×10^{-6}	0.0004	0.0008	0.0018
D ₁₀	0.70	4.4×10^5	6.6×10^{-1}	1.00×10^0	1.3129	1.1197	0.9550	D ₁₀	0.62	1.9×10^7	2.9×10^1	1.00×10^0	1.1080	0.9739	0.8561
D ₅	0.76	1.6×10^5	2.4×10^{-1}	9.79×10^{-1}	1.1511	1.1452	1.1391	D ₅	0.52	2.1×10^8	3.1×10^2	9.86×10^{-1}	1.7464	1.2158	0.8465
D ₀	0.86	3.5×10^4	5.3×10^{-2}	5.90×10^{-1}	0.8537	1.0485	1.2871	D ₀	0.66	4.3×10^6	6.5×10^0	6.81×10^{-1}	1.1313	1.0779	1.0269

^aThe biasing potentials considered are illustrated in Figure 3. The parameters α_n , $\tilde{\tau}_n$, F_n , and $\tilde{C}_n(t_b)$ are defined by eqs 6–10, where $\tilde{C}_n(t_b)$ is the ratio of $C_n(t_b)$ for a given simulation to the corresponding value for the unbiased simulation U , and reference times $t_b = 1, 10, \text{ or } 100$ ns are considered. The results are also illustrated graphically in Figure 5.

However, in most simulations, the effects of the biasing potentials D_{10} and D_5 are comparable, and the corresponding enhancement is marginal (for Ala₄ and Val₄, the searching is even slightly slower than that of plain MD). Only the biasing potential D_0 enhances the searching nearly systematically (with the possible exception of Val₄). Considering all systems, the numbers of conformations that have been visited within 50 ns changes by factors of 0.7–1.2 (D_{10}), 0.8–1.1 (D_5), and 0.9–1.9 (D_0) compared to the unbiased simulation U . For the biasing potential D_0 , this factor tendentially increases with the oligopeptide length n and is typically larger for Ala _{n} compared with Val _{n} at a given n .

The different $N_n(t)$ curves of Figure 4 could be adequately fitted to stretched-exponential functions (eqs 6 and 7). The values of the resulting parameters α_n (stretching exponent), τ_n (characteristic searching time), and $\tilde{\tau}_n$ (effective visiting time) for the different Ala _{n} and Val _{n} oligopeptide simulations are reported in Table 1 and displayed graphically in Figure 5a,b. The α_n values range from 0.22 (high revisiting tendency) to 1.00 (random configuration generation process), the τ_n values span 9 orders of magnitude from 0.26 ns to 0.21 s, and the $\tilde{\tau}_n$ values span 6 orders of magnitude from 0.34 ps to 0.31 μ s for the different systems and biasing potentials considered.

The α_n values associated with the unbiased MD simulations U range from 0.22 to 0.77. They systematically increase with n for a given oligopeptide type and are systematically higher for Ala _{n} compared with Val _{n} at a given n . This indicates a high revisiting tendency, expectedly more pronounced for systems of lower dimensionality (low n , implying a higher probability to diffuse back into previously visited conformations) and systems involving stronger non-local interactions (more important hydrophobic side chain–side chain interactions in Val _{n} compared with Ala _{n} , inducing a more significant conformational probability bias). The τ_n values for the unbiased simulations increase roughly exponentially with the oligopeptide length n , approximately by one order of magnitude upon increasing n by two for both Ala _{n} and Val _{n} . The values are also systematically higher by about one order of magnitude for Val _{n} compared with Ala _{n} . Thus, for example, a coverage of a fraction $1 - e^{-1}$ (63%) of the entire conformational space accessible to the Ala₁₀ and Val₁₀ oligopeptides can be estimated to require plain MD simulations on the 0.2 and 6 ms time scales, respectively. These estimates should probably be regarded as lower bounds, because they rely on α_n and τ_n values characteristic of the (low free energy) regions visited on the 50 ns time scale. In reality, the conformational probability bias is likely to increase when the sampling is extended to other (higher free energy) regions, probably resulting in a decrease of the α_n value and an increase of the τ_n value (or even a breakdown of the stretched-exponential fitting). Note also that these time scales are far above the suggested time scales for secondary-structure formation in polypeptides (section 1), indicating that these processes are by no means random search processes and require the dynamical sampling of a much smaller volume fraction of conformational space (e.g., compared with $1 - e^{-1}$) for their occurrence (Levinthal's paradox⁸⁸). Finally, the $\tilde{\tau}_n$ values for the unbiased simulations

evidence much smaller variations (compared to τ_n) across the different systems considered, ranging from 0.29 to 1.65 ns for Ala _{n} and from 6.32 to 22.8 ns for Val _{n} . These times nearly systematically decrease with n for a given oligopeptide type and are about one order of magnitude higher for Val _{n} compared with Ala _{n} at a given n . This suggests effective times associated with backbone torsional angle transitions (between wells as defined in section 2.3) on the order of 1 and 10 ns for the two types of peptides. The effective times $\tilde{\tau}_n$ are expectedly lower for systems of higher dimensionality (high n , because a conformational transition results from a transition in any of the n linkages) and higher for systems involving more important transition barriers (more bulky side chains in Val _{n} compared with Ala _{n} , inducing higher inter-conversion barriers; Figure 2).

The α_n values for the simulations with the BF biasing potentials range from 0.41 to 1.00. For a given oligopeptide, they nearly systematically increase along the series F_{10} , F_{20} , and F_{30} and are nearly systematically higher than the corresponding value for the unbiased simulation U . Here also, the values tendentially increase with n for a given oligopeptide type (the comparison of the Ala _{n} to the corresponding Val _{n} oligopeptide does not reveal systematic trends). For some systems (Ala₄, Ala₁₀, Val₈, and Val₁₀), α_n may become very close to one, suggesting that the FB-LEUS scheme leads in this case to a trajectory that is very similar to a random configuration generation process. Simultaneously, the τ_n and $\tilde{\tau}_n$ values are dramatically decreased compared to the plain MD simulation, by factors of about 25–250 (F_{10}), 200–1700 (F_{20}), or 500–7500 (F_{30}) for the different systems. For example, the coverage of a fraction $1 - e^{-1}$ (63%) of the entire conformational space accessible to the Ala₁₀ or Val₁₀ oligopeptides using the F_{30} biasing potential can be estimated to require simulations on the 200 and 800 ns time scales, respectively. These estimates are probably more realistic than the corresponding estimates (0.2 and 6 ms, respectively) for the unbiased simulation U (see above), considering that the F_{30} biasing potential largely reduces the conformational probability bias in the simulated ensemble (at least the local single-linkage component of this bias). The $\tilde{\tau}_n$ values are in the ranges 6.7–32 ps (F_{10}), 0.78–4.3 ps (F_{20}), and 0.34–1.6 ps (F_{30}) for Ala _{n} , and 32–230 ps (F_{10}), 6.5–13 ps (F_{20}), and 1.1–6.1 ps (F_{30}) for Val _{n} . Here also, for each of the three biasing potentials, these times decrease with n for a given oligopeptide type and are shorter for Ala _{n} compared with Val _{n} at a given n . The effective time associated with backbone torsional angle transitions has thus been brought from the 1–10 ns range for the unbiased simulations to the 1–100 ps range for the biased simulations.

The α_n values for the simulations with the VD biasing potentials range from 0.20 to 0.86. For a given oligopeptide, the values for the series D_{10} , D_5 , and D_0 are generally similar, and comparable to the corresponding value for the unbiased simulation U . Here also, the values tendentially increase with n for a given oligopeptide type (the comparison of the Ala _{n} to the corresponding Val _{n} oligopeptide does not reveal systematic trends). Similarly, the τ_n and $\tilde{\tau}_n$ values are only moderately altered and in a nonsystematic way compared to the plain MD simulation. Note, however, that the values for

the biasing potential D_0 are nearly systematically lower than those for the unbiased simulation (except for Val₄ and Val₆), in agreement with the searching enhancement generally observed for this scheme (Figure 4).

The results of the simulations in terms of the statistical efficiency parameter F_n (eq 8) for the different Ala_{*n*} and Val_{*n*} oligopeptide simulations are reported in Table 1 and displayed graphically in Figure 5c. In terms of statistical efficiency, the application of the FB-LEUS biasing potentials may lead to a dramatic deterioration compared to the unbiased MD simulation U .

Because plain (thermostatted) MD samples in the canonical ensemble, the corresponding value of F_n is one for all n (maximal statistical efficiency). The value corresponding to a simulation where a single configuration entirely dominates the reweighted probability distribution is $F_n = N_f^{-1} = 4 \times 10^{-8} \approx 0$ (minimal statistical efficiency). For the different oligopeptides considered, the values of F_n for the biased simulations range from 8.0×10^{-6} to 1.0. Values lower than one indicate a tendency to generate irrelevant configurations in terms of the conformational distribution characteristic of the physical system, that is, high-energy configurations with low Boltzmann weights in the physical ensemble.

The deterioration of the statistical efficiency is most pronounced when using the BF biasing potentials, the efficiency systematically decreasing along the series F_{10} , F_{20} , and F_{30} of increasingly “flattened” mono-peptide free-energy surfaces (Figure 3a,b). For example, considering Ala₄ with the biasing potential F_{30} , only about 0.4% of the trajectory configurations actually contribute to the statistics relevant for the physical ensemble, the corresponding fraction becoming as low as 0.001% for Ala₁₀. The deterioration is much less important when using the VD biasing potentials, the efficiency systematically decreasing along the series D_{10} , D_5 , and D_0 of increasingly “dug” valleys on the mono-peptide free-energy surfaces (Figure 3c,d). For example, considering all systems, at least 97% of the trajectory configurations contribute to the relevant statistics for the biasing potentials D_{10} and D_5 , while this fraction is still at least 59% for the biasing potential D_0 . This large qualitative difference between the BF and VD biasing potentials is easily understood on the basis of the very different extents of irrelevant conformational regions they open up to sampling (large basins vs narrow valleys; Figure 3). The trends observed within the two series of potentials are easily rationalized using similar arguments.

The value of F_n systematically decreases with n for a given oligopeptide and biasing potential type (the comparison of the Ala_{*n*} to the corresponding Val_{*n*} oligopeptide reveals comparable values, but no further systematic trends). This decrease results from two main effects. First, distinguishing between relevant and irrelevant conformational regions at the mono-peptide level (in terms of the distribution appropriate for the physical system), the probability that all peptide bonds are simultaneously found in relevant regions at the oligopeptide level decreases exponentially with n (combinatorial effect). Second, the extension of the peptide length increases the likelihood that the location of the free-energy minima at the

oligopeptide level differ from the corresponding locations at the mono-peptide level, as a consequence nonlocal interactions (i.e., beyond the local conformational preferences of the individual linkages), resulting in a tendency of the FB-LEUS biasing potential to drive the system toward conformational regions that do not correspond to free-energy basins at the oligopeptide level.

As illustrated in Figure 6, the combinatorial effect is dominant in the context of the BF biasing potentials. In this case, F_n decreases as a single exponential function of n (no prefactor), that is, as $F_n \approx 10^{-an}$, where the value of a is nearly identical for Ala_{*n*} and Val_{*n*}, namely, about -0.29 , -0.49 , and -0.54 for F_{10} , F_{20} , and F_{30} , respectively. The existence of such a relationship suggests in particular that a good estimate for F_n at the oligopeptide level can already be formulated from the knowledge of the statistical efficiency F_1 at the mono-peptide level. Note also that a does not decrease linearly with the free-energy cutoff h used in the definition of the BF potential (section 2.2 and Figure 3a). The dependence of a on h is expected to level off at a value h_{\max} equal to the minimum-to-maximum difference in the fragment free-energy map. For any h above h_{\max} , the most statistically inefficient BF potential has been reached (entirely homogeneous sampling of the relevant conformational subspace at the fragment level). Considering Figure 3b, this limit is essentially reached for the biasing potential F_{30} (at least for poly-Ala) and one has thus $F_n \approx 10^{-0.54n}$ for any $h \geq 30$ kJ mol⁻¹.

The opposing effects of the enhancement in the searching efficiency and of the deterioration in the statistical efficiency can be characterized by means of the combined sampling efficiency parameter $C_n(t_0)$ associated with a given time scale t_0 (eq 10). The results of the different Ala_{*n*} and Val_{*n*} oligopeptide simulations in terms of the corresponding relative values $\tilde{C}_n(t_0)$, with unbiased simulations U taken as reference, are reported in Table 1 and displayed graphically in Figure 5d,e,f for $t_0 = 1, 10, \text{ or } 100$ ns. The values of $\tilde{C}_n(t_0)$ for the unbiased simulations are all equal to one by definition. Most of the biased simulations present values lower than one, indicating a decrease of the combined sampling efficiency compared with plain MD. Exceptions involve the BF biasing potentials for Ala₄ and $t_0 = 100$ ns, as well as the VD biasing potentials in about half of the considered systems and t_0 combinations. However, even in these cases, the sampling enhancement never exceeds a factor two.

For the simulations relying on BF biasing potentials, $\tilde{C}_n(t_0)$ tendentially decreases upon comparing Val_{*n*} to Ala_{*n*}, along the series F_{10} , F_{20} , and F_{30} , upon increasing n , or upon decreasing t_0 . The results suggest that the BF scheme may become competitive with plain MD for the systems considered when the simulation time scale t_0 is of the same order as the effective searching time τ_n of the unbiased simulation (e.g., $t_0 = 100$ ns compared with $\tau_n = 134$ ns for Ala₄, where the BF schemes are more efficient than plain MD). This is of course a disappointing conclusion since this time scale is precisely the one beyond which a sampling enhancement is in principle no longer needed.

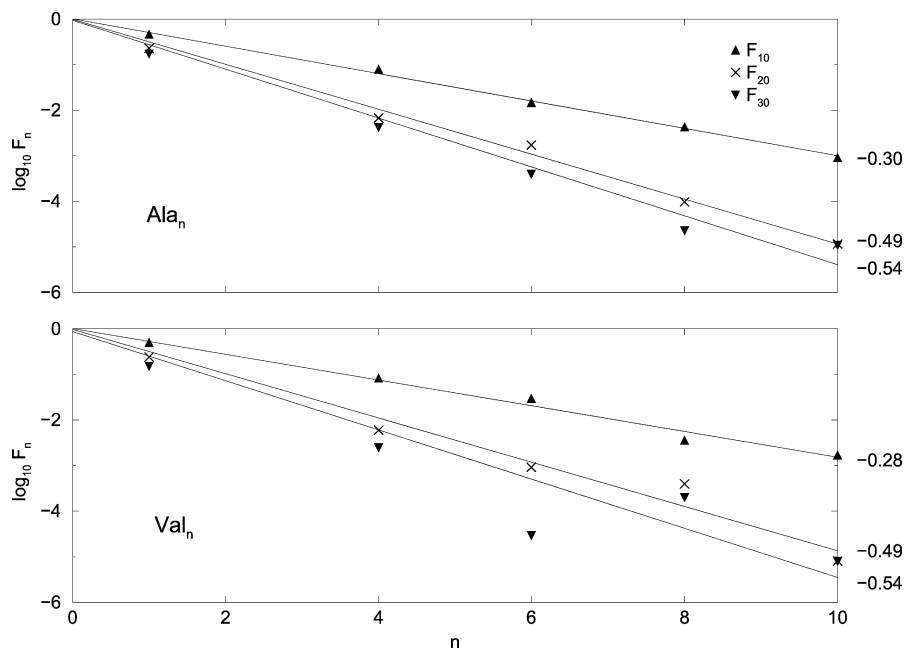


Figure 6. Statistical efficiency F_n for Ala_n and Val_n oligopeptide simulations performed with different FB-LEUS biasing potentials of the BF type. The BF biasing potentials considered are illustrated in Figure 3a,b. The parameter F_n is defined by eq 8. It is displayed on a (decimal) logarithmic scale, along with least-squares-fit regression lines anchored at the origin (single exponential fit, no prefactor). The corresponding slopes are also indicated. Corresponding numerical values can be found in Table 1 except for $n = 1$ ($\text{Ala}_1 = 0.47, 0.23, 0.17$, and $\text{Val}_1 = 0.51, 0.24, 0.15$, for biasing potentials F_{10}, F_{20} and, F_{30} , based on the N&C-blocked mono-peptides).

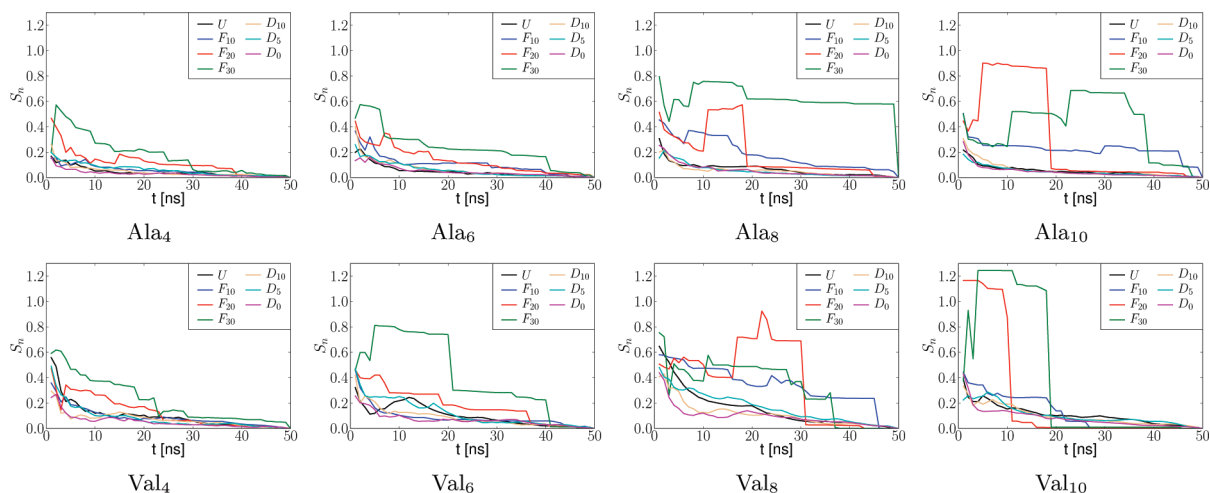


Figure 7. Free-energy convergence score $S_n(t)$ for Ala_n and Val_n oligopeptide simulations performed with different FB-LEUS biasing potentials. The biasing potentials considered are illustrated in Figure 3. The score $S_n(t)$ is defined by eq 11.

For the simulations relying on the VD biasing potentials, $\tilde{C}_n(t_0)$ is typically close to one, the dependence on the type of peptide and choice of biasing potential, as well as on n and t_0 , being rather nonsystematic. Note, however, that $\tilde{C}_n(t_0)$ is nearly always higher than one when the biasing potential D_0 is used (except for Val_4 and Val_8 , as well as Ala_{10} with $t_0 = 1$ ns), with values ranging from 0.72 to 1.72. In other words, this biasing potential generally leads to a modest sampling enhancement at all time scales considered.

The above results clearly illustrate the problem of the trade-off between searching and statistical efficiencies as determinants of the overall sampling efficiency of a scheme. For practical applications, a more directly relevant point to be addressed concerns the reliability with which the low free

energy conformations and associated free energies can be determined using a given sampling scheme (convergence with time, number of interconversion transitions, and mutual overlap across schemes). These properties are investigated in turn below.

The results of the simulations in terms of the score $S_n(t)$ measuring the free-energy convergence (eq 11) for the different Ala_n and Val_n oligopeptide simulations are displayed in Figure 7. The score $S_n(t)$ accounts for the root-sum-square error in the predicted conformational populations for a simulation of duration t compared with the full simulation of duration t_s (50 ns). All the $S_n(t)$ curves start somewhat below one for $t \approx 0$ and converge to exactly zero at $t = t_s$. However, the rapidity as well as

the degree of monotonicity and smoothness of this evolution provides information concerning the convergence of the predicted lowest free energy conformations and of the associated free energies. For example, a rapid variation at a given time t suggests that the sudden visit to a new low free energy conformation has radically altered the population distribution within the most populated states of the (reweighted) ensemble, that is, that the distribution just before time t was unconverged. The curves corresponding to the unbiased simulations are rapidly converging as well as essentially smooth and monotonic. The same applies to the curves corresponding to the simulations relying on the VD biasing potentials. A convergence improvement is visible in some systems (e.g., Val₄ and Val₈) for some of these potentials. In contrast, the curves corresponding to the simulations relying on the BF biasing potentials are erratic. This behavior becomes increasingly pronounced upon increasing the oligopeptide length n . In other words, the predicted most stable conformations and associated free energies are steadily reshuffled by the sporadic encounter of new low free energy conformations and certainly not converged after 50 ns of simulation (with the possible exception of Ala₄ and Val₄).

The results of the simulations in terms of the average number of (direct or indirect) interconversion transitions $T_n(G_c)$ between the lowest free energy conformations (section 2.3) for the different Ala _{n} and Val _{n} oligopeptide simulations are displayed in Figure 8, considering three free-energy cutoffs $G_c = 5, 10, \text{ or } 15 \text{ kJ mol}^{-1}$. These numbers are correlated with the accuracy with which the relative free energies of the lowest free energy conformations will be predicted by a given scheme, because a sufficient number of interconversion transitions is a prerequisite for the evaluation of an accurate free-energy difference. For the cutoff values considered, the average number of transitions to each low free energy conformation is on the order of $10^2\text{--}10^4$ for the unbiased simulations and the simulations relying on the VD biasing potentials. The $T_n(G_c)$ values tendentially decrease upon increasing the oligopeptide length n and upon increasing G_c , and are systematically slightly higher for the simulations employing the biasing potential D_0 compared to the corresponding unbiased simulations (increase by factors 1.1–5.9 for the different systems and G_c values considered). In contrast, for the simulations relying on the BF biasing potentials, the $T_n(G_c)$ values decrease much more abruptly upon increasing the peptide length, as well as along the series F_{10} , F_{20} , and F_{30} . As a result, if the numbers of transitions are comparable to those of the unbiased and VD simulations for Ala₄ and Val₄, they may decrease to very few (or even a single one) for some of the Ala₁₀ and Val₁₀ simulations. For these simulations, the ranking of the low free energy conformations and the corresponding relative free-energy estimates are likely to be incorrect. A second observation is that, in contrast to the unbiased and VD simulations, the $T_n(G_c)$ values for the BF simulations sometimes increase upon increasing G_c . For the unbiased and VD simulations, a systematic

decrease of $T_n(G_c)$ upon increasing G_c is expected, because it extends the averaging to less populated conformations involved in fewer transitions. For the BF simulations, nonsystematic changes result from the fact that, especially for high n and low G_c , the number of states $L_n(G_c)$ below G_c is very low (poor statistics).

The results of the simulations in terms of the overlaps $O_n^{AB}(G_c)$ and $O_n^{BA}(G_c)$, between the lowest free energy conformations generated by all pairs (A and B) of schemes (section 2.3) for the different Ala _{n} and Val _{n} oligopeptide simulations are displayed in Figure 9, considering three free-energy cutoffs $G_c = 5, 10, \text{ or } 15 \text{ kJ mol}^{-1}$. Expectedly, the extent of overlap between the low free energy conformations predicted by the different schemes decreases upon increasing n . However, this decrease is much more pronounced for the schemes relying on the BF biasing potentials and, among these, in the sequence F_{10} , F_{20} , and F_{30} . For Ala₄ and Val₄, all schemes present a very high extent of mutual overlap. However, for Ala₁₀ and Val₁₀, only simulations U , D_{10} , D_5 , and D_0 , and to a lesser extent F_{10} , present a reasonable extent of mutual overlap. In contrast, simulations F_{20} and F_{30} have produced two different sets of low free energy conformations, presenting essentially no overlap with the latter sets and with each other.

Similar observations can be made based on the ranking $R_{m,i}$ and free energies $G_{m,i}$ attributed by a given scheme i to the five conformations m with the lowest consensus rank R_m (section 2.3). The values of $R_{m,i}$ and $G_{m,i}$ for the different schemes are reported in Table 2. The conformations m associated with the lowest R_m are systematically found to be those with $m = 0$, corresponding (via their binary codes) to the n successive dihedral-angle pairs (ϕ_i, ψ_i) within the lowest free energy well at the mono-peptide level (Figure 2). Bundles of 20 illustrative trajectory configurations assigned to these conformations are represented in Figure 10. The four other conformations with lowest consensus ranks differ from these ones by one or two bits only (with reference to the corresponding binary codes), commonly the first or last ones or both, indicating a different free-energy well at the mono-peptide level for dihedral-angle pairs (ϕ_1, ψ_1) or (ϕ_n, ψ_n) . This observation suggests that nonlocal interactions (beyond single-linkage conformational preferences) are weak in the oligopeptides considered (and given their relatively short lengths). This is confirmed by a DSSP analysis⁸⁹ (data not shown) of the corresponding configurations, which suggests a quasi-total absence of secondary structure (beyond turns and bends, that is, no helices or sheets). For Ala₄ and Val₄, nearly all schemes have predicted conformation 0 as the lowest free energy one (except simulation F_{30} for Val₄, for which this conformation has rank 2). Most simulations also encompass a significant fraction of the four other conformations among their lowest free energy conformations. For Ala _{n} and Val _{n} with $n = 6, 8, \text{ or } 10$, the plain MD simulations and the simulations relying on the VD biasing potentials have all predicted conformation 0 as the lowest free energy one. However, the agreement concerning the ranking and relative free energies of the four other conformations progressively worsens upon increasing n . In contrast,

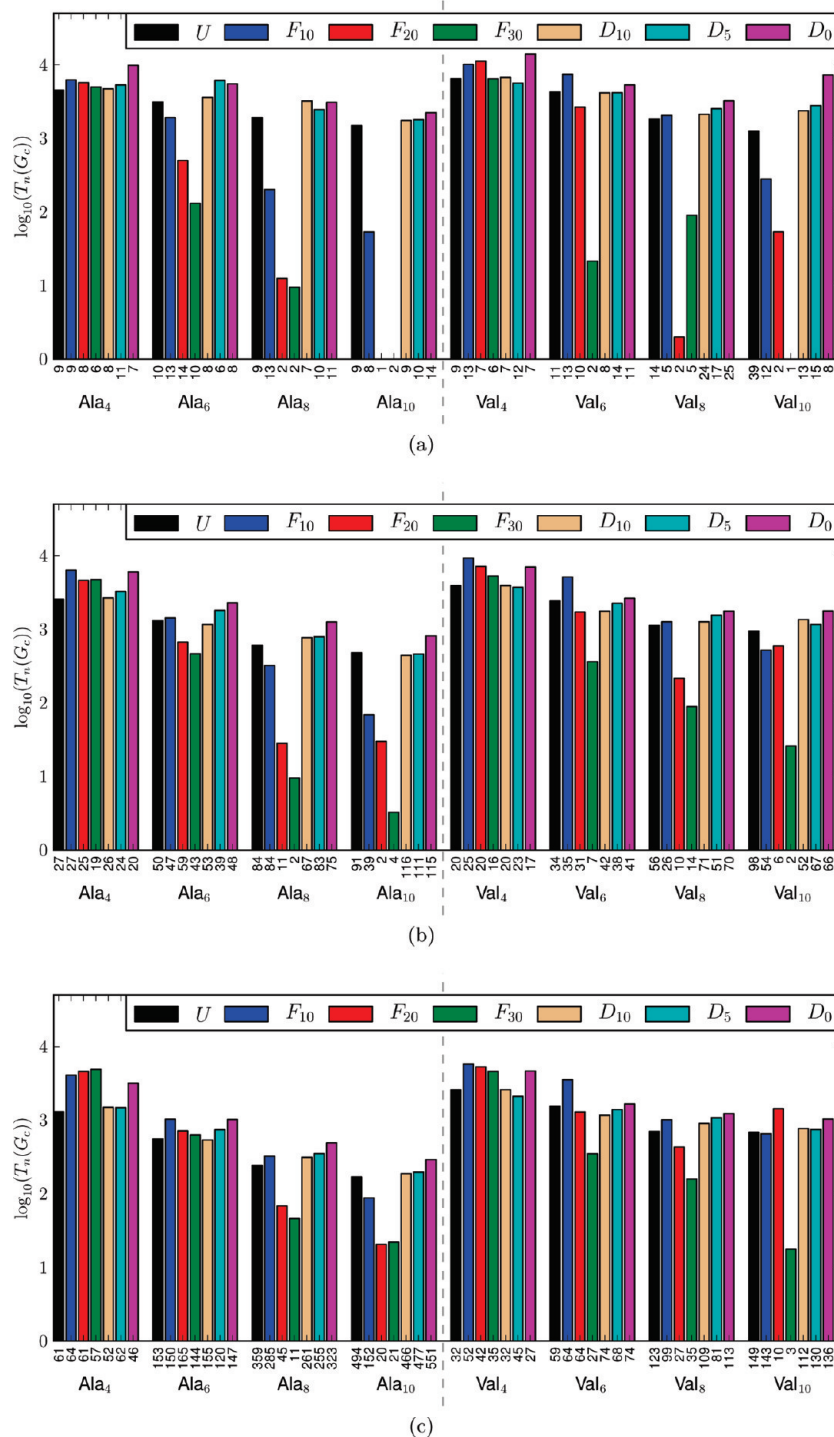


Figure 8. Average number of (direct or indirect) transitions $T_n(G_c)$ between the $L_n(G_c)$ lowest free energy conformations for Ala_{*n*} and Val_{*n*} simulations performed with different FB-LEUS biasing potentials. The biasing potentials considered are illustrated in Figure 3. The quantities $L_n(G_c)$, reported below the individual bars, and $T_n(G_c)$ are defined in section 2.3. The values are calculated using free-energy cutoffs $G_c = 5, 10, \text{ or } 15 \text{ kJ mol}^{-1}$. Note the use of a (decimal) logarithmic scale.

except for Ala₆ with biasing potentials F_{10} and F_{20} , the simulations relying on the BF biasing potentials fail to produce the same lowest free energy conformation and to agree with each other concerning this conformation beyond $n = 4$. In other words, the biased ensemble is in this case crowded with irrelevant configurations, leading to very high uncertainties in the calculated relative free energies for the few relevant (low free energy) conformations and, for longer

peptides, to the complete omission of most of these conformations.

4. Conclusion

The goal of the present study was to expand the scope of the LEUS method to solvated polymers with more than a few relevant degrees of freedom, by means of a fragment-

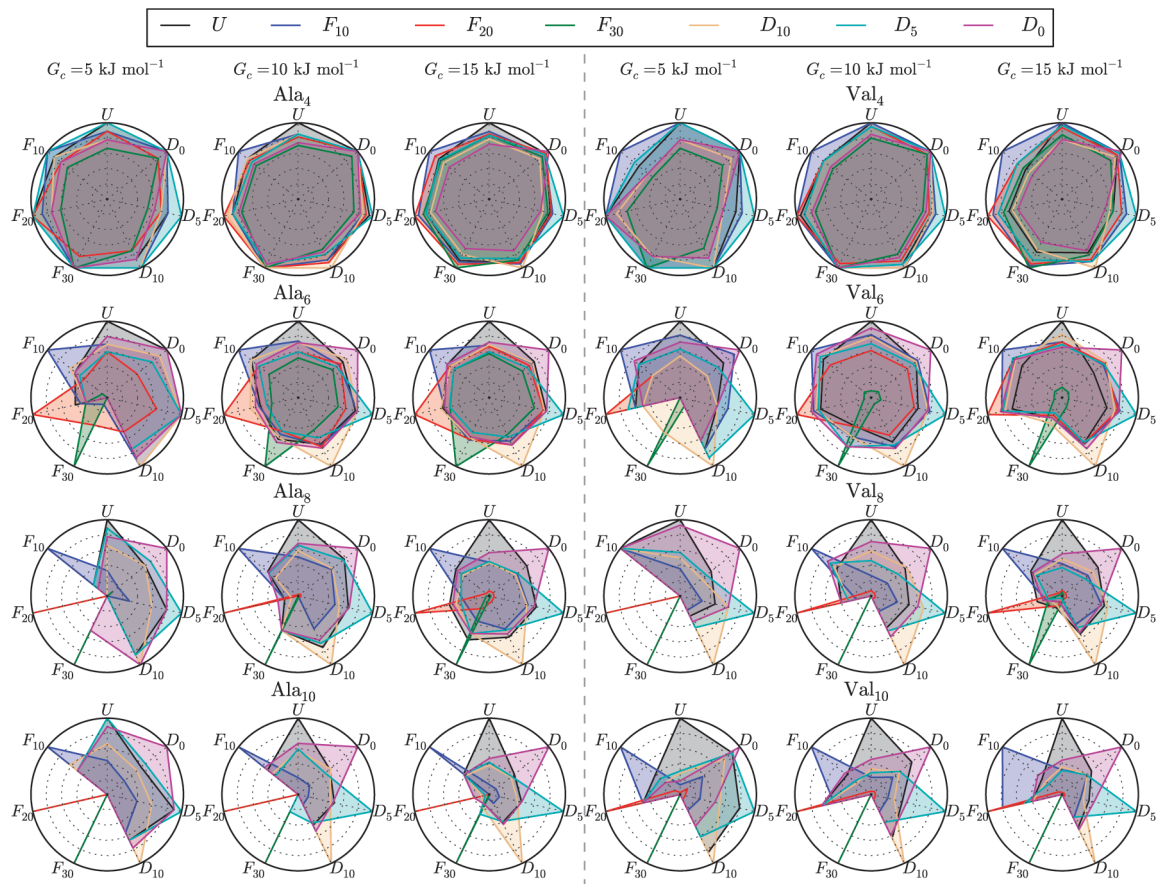


Figure 9. Overlaps $O_n^{AB}(G_c)$ and $O_n^{BA}(G_c)$ between the lowest free energy conformations predicted by pairs (A and B) of simulations of the Ala_n and Val_n oligopeptides performed using different FB-LEUS biasing potentials. The biasing potentials considered are illustrated in Figure 3. The quantities $O_n^{AB}(G_c)$ and $O_n^{BA}(G_c)$ are defined in section 2.3. Radial labels indicate the scheme A and line colors the scheme B. The intercept between a given radial line (A) and a given colored line (B) represents the overlap, $O_n^{AB}(G_c)$. The overlap $O_n^{BA}(G_c)$ can be read from the inverted pair of radial and colored lines. The origin of the circle corresponds to an overlap of zero, while its perimeter corresponds to an overlap of one. The values are calculated using free-energy cutoffs $G_c = 5, 10, \text{ or } 15 \text{ kJ mol}^{-1}$.

based approach. The resulting scheme, FB-LEUS, can be applied along with basin-filling (BF) or valley-digging (VD) fragment-based biasing potentials. The feasibility of this scheme was tested in the context of (unblocked) polyalanine (poly-Ala) and polyvaline (poly-Val) oligopeptides, using fragment-based biasing potentials optimized at the (partially) blocked monopeptide level. The results show that the application of the FB-LEUS biasing potentials may lead to an impressive enhancement of the searching power (volume of conformational space visited in a given amount of simulation time). However, this enhancement is largely offset by a deterioration of the statistical efficiency (representativeness of the biased ensemble in terms of the conformational distribution appropriate for the physical ensemble). As a result, it appears difficult to engineer FB-LEUS schemes representing a significant improvement over plain MD, at least for the systems considered here.

This no-free-lunch conclusion might seem disappointing at first sight, and it is interesting to analyze in more detail the reasons for this failure. The problems encountered by any sampling-enhancement method relying on a memory-based biasing potential, such as the LEUS approach, in the context of a relevant conformational subspace of high dimensionality can be summarized as follows:

- Internal Dimensionality Problem.** The dimensionality of the subspace involved in the construction of the biasing potential, referred to here as its internal dimensionality (as opposed to the true dimensionality, that is, that of the relevant conformational subspace), cannot be too high due to memory costs and build-up duration requirements. This problem is addressed in the FB-LEUS scheme by using a fragment-based approach, where the internal dimensionality N_{LE} refers to the fragments and the true dimensionality $N_{\text{US}} = N_{\text{F}}N_{\text{LE}}$ to the real system encompassing N_{F} fragments.
- Irrelevant Volume Problem.** In order to enhance the rate of discovery of new relevant conformational states (i.e., the conformational-searching efficiency), as well as the statistics concerning the relative populations of these states (i.e., on their relative free energies), the biasing potential must facilitate interconversion transitions between states. This implies the creation of low free energy pathways connecting these states and, thus, the opening of an irrelevant volume of conformational space to sampling, irrelevant meaning that the accessed configurations are characterized by very low Boltzmann weights in the physical ensemble. The size of this irrelevant volume is the main determinant of the

Table 2. Free-Energy Ranking $R_{m,i}$ and Relative Free Energy $G_{m,i}$ (in parentheses, in kJ mol^{-1}) of the Five Consensus Lowest Free Energy Conformations m (Lowest Consensus Rank R_m) for Ala_n and Val_n Simulations Performed with the Seven Different FB-LEUS Biasing Potentials^a

m	Δ_m	U	F_{10}	F_{20}	F_{30}	D_{10}	D_5	D_0	m	Δ_m	U	F_{10}	F_{20}	F_{30}	D_{10}	D_5	D_0
1	0	1(0.0)	1(0.0)	Ala ₄ 1(0.0)	1(0.0)	1(0.0)	1(0.0)	1(0.0)	1	0	1(0.0)	1(0.0)	Val ₄ 1(0.0)	2(0.7)	1(0.0)	1(0.0)	2(0.3)
2	64	2(1.2)	2(1.6)	2(0.7)	6(3.7)	2(1.2)	2(0.8)	2(1.2)	2	64	2(1.3)	2(0.5)	2(0.8)	4(2.8)	3(1.5)	2(0.1)	1(0.0)
3	128	3(1.5)	3(1.9)	3(1.4)	2(2.4)	3(1.6)	3(1.6)	3(1.5)	3	128	3(1.4)	3(1.3)	3(1.5)	1(0.0)	2(1.3)	3(1.5)	3(1.5)
4	1	5(2.4)	4(2.6)	5(3.4)	3(2.6)	4(2.2)	4(2.2)	4(2.3)	4	1	4(2.0)	4(1.8)	4(2.8)	3(2.3)	4(2.3)	4(1.8)	5(2.8)
5	16	4(2.3)	5(3.7)	7(3.9)	11(6.1)	5(3.9)	5(3.1)	7(4.1)	5	65	6(3.6)	5(2.1)	6(3.2)	9(6.0)	5(3.9)	5(2.6)	4(2.8)
1	0	1(0.0)	1(0.0)	Ala ₆ 1(0.0)	27(8.1)	1(0.0)	1(0.0)	1(0.0)	1	0	1(0.0)	2(0.0)	4(1.8)	6(9.7)	1(0.0)	1(0.0)	1(0.0)
2	1024	2(0.9)	2(1.5)	9(4.3)	24(7.8)	2(1.6)	2(1.5)	3(1.3)	2	1024	2(0.7)	1(0.0)	1(0.0)	3(7.6)	2(1.6)	2(0.7)	2(1.1)
3	2048	3(1.6)	3(1.7)	15(5.0)	16(6.1)	3(1.7)	3(1.5)	2(1.2)	3	2048	3(1.5)	3(1.4)	6(2.3)	110(22.5)	3(2.1)	3(1.0)	3(1.2)
4	1	4(2.4)	4(2.6)	2(1.6)	18(6.5)	4(2.4)	4(2.7)	4(2.7)	4	1	4(2.2)	4(1.9)	8(3.2)	83(20.9)	4(3.1)	6(2.6)	4(2.4)
5	1040	16(6.1)	10(4.5)	3(3.0)	3(0.3)	17(6.7)	15(7.4)	10(5.2)	5	64	144(29.6)	16(5.7)	2(0.6)	97(21.7)	6(4.2)	4(1.2)	13(5.5)
1	0	1(0.0)	25(6.5)	Ala ₈ 1283(31.5)	2307(41.9)	1(0.0)	1(0.0)	1(0.0)	1	0	1(0.0)	2(2.3)	35(16.5)	101(22.2)	1(0.0)	1(0.0)	1(0.0)
2	16384	2(1.2)	71(9.3)	409(24.8)	5077(49.3)	2(1.3)	2(1.3)	3(2.0)	2	16384	2(1.5)	37(10.8)	13(11.1)	16(10.2)	3(0.7)	2(0.5)	2(0.6)
3	32768	3(1.6)	26(6.5)	511(26.1)	2832(43.7)	3(2.1)	3(1.4)	2(1.7)	3	32768	3(1.5)	5(4.5)	4(8.0)	766(38.2)	7(1.8)	3(1.2)	4(1.0)
4	1	4(2.9)	33(6.9)	1415(32.2)	561(31.6)	4(2.6)	4(2.5)	4(2.2)	4	1	4(2.1)	3(2.5)	138(24.1)	41(17.3)	6(1.5)	8(3.2)	7(2.2)
5	16	18(6.1)	321(15.6)	43(14.6)	2(0.4)	11(5.8)	13(5.1)	9(4.2)	5	32769	9(3.9)	1(0.0)	99(22.0)	396(32.2)	9(2.8)	16(4.8)	14(4.1)
1	0	1(0.0)	61(11.2)	Ala ₁₀ 4422(43.9)	1111(38.0)	1(0.0)	1(0.0)	1(0.0)	1	0	1(0.0)	7(3.3)	3(6.4)	13920(94.3)	1(0.0)	1(0.0)	2(0.3)
2	524288	3(1.8)	14(6.4)	106(22.7)	9350(55.8)	2(1.5)	2(1.1)	2(1.1)	2	262144	4(0.2)	6(3.1)	4(6.6)	X	2(1.8)	2(0.1)	1(0.0)
3	262144	2	8(5.0)	2020(38.1)	11245(57.8)	3(1.8)	3(1.7)	3(1.2)	3	524288	5(0.9)	3(0.5)	2(4.8)	X	3(1.8)	3(1.5)	3(1.7)
4	1	4(2.5)	3(2.5)	59129(75.9)	4032(47.9)	4(2.5)	4(3.0)	4(1.8)	4	1	14(2.2)	17(5.9)	22(19.0)	5027(78.5)	4(2.5)	4(2.5)	4(2.3)
5	262145	2,20	6(4.0)	28318(63.2)	33176(71.6)	5(3.8)	8(4.5)	6(3.4)	5	4096	245(24.3)	1(0.0)	11113(88.0)	3514(74.0)	295(30.5)	13(4.8)	11(5.6)

^a The biasing potentials considered are illustrated in Figure 3. The quantities m (integer corresponding to the binary code of a unique peptide backbone conformation), R_m , and $G_{m,i}$ are defined in section 2.3. The column Δ_m lists the bits of the binary code represented by m ($2n$ bits in total) that differ from zero. An "X" indicates that a conformation was never visited.

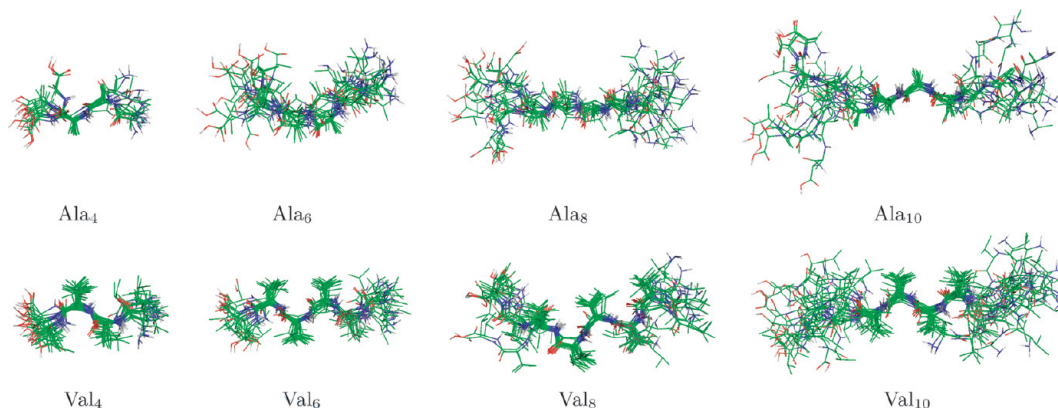


Figure 10. Illustrative configurations assigned to the backbone conformation $m = 0$ with the lowest consensus rank R_m for Ala_n and Val_n simulations performed with the seven different FB-LEUS biasing potentials. The biasing potentials considered are illustrated in Figure 3. The quantities m (integer corresponding to the binary code of a unique peptide backbone conformation) and R_m are defined in section 2.3. For each system, a bundle of 20 representative structures was extracted from successive 0.5 ns blocks along the first 10 ns of the simulation U . The structures were superimposed by rototranslational least-squares-fitting based on all backbone atoms of the four central residues (for Ala_4 the two central residues).

statistical efficiency of a biasing approach. In the context of the FB-LEUS scheme, the limitation of the irrelevant volume at the fragment level is particularly important because the corresponding volume for the real system increases exponentially with the number N_F of fragments in this system (combinatorial effect). In this respect, VD-type biasing potentials appear far superior to BF-type biasing potentials, because they promote transitions between states while opening a minimal amount of irrelevant volume to sampling. However, the oligopeptide test systems considered here do not present very high transition barriers (poly-Val slightly more than poly-Ala), so that transition rates are only moderately increased by the VD scheme.

The additional problems encountered specifically by the FB-LEUS approach can be summarized as follows:

C. Free-Energy Additivity Approximation. The LEUS scheme in terms of BF-type biasing potentials relies on the principle that a biasing potential that is approximately equal to the negative of the free-energy hypersurface in the relevant conformational subspace up to a given threshold above the global minimum, will lead to a nearly homogeneous coverage of this region in the biased simulation. However, a FB-LEUS biasing potential will only satisfy this condition if the free-energy function of the real system can be expressed as a sum of independent contributions from the N_F fragments. In reality, this sum can only account for local (single-fragment) contributions, thereby neglecting nonlocal (fragment-coupling) effects. Similar considerations apply to the VD-type biasing potentials in terms of the locations of the free-energy minima and magnitudes of transition barriers at the fragment and real-system levels. The presence of nonlocal effects has both negative and positive consequences for the FB-LEUS scheme. On the negative side, the FB-LEUS biasing potential is not strictly appropriate for the homogeneous (BF) or accelerated (VD) sampling of the real system, which may lead to a decrease in its searching efficiency. On the positive side, the presence

of residual nonlocal driving forces may steer the biased system away from irrelevant conformational regions (e.g., by secondary structure formation or folding), thereby increasing its statistical efficiency. However, the oligopeptide test systems considered here do not present strong nonlocal interactions (poly-Val slightly more than poly-Ala) and appear to be significantly affected by neither of the two effects.

D. Combinatorial Problem. In cases where nonlocal effects are negligible, the application of the FB-LEUS scheme with a BF-type biasing potential will promote facilitated interconversions between a finite number M of equally deep free-energy basins at the fragment level (e.g., $M = 4$ for the present oligopeptide systems). Thus, it will induce the sampling of M^{N_F} equally populated states for a real system encompassing N_F fragments. Even if the searching rate is increased, visiting as many states, if tractable at all, requires a sizable amount of simulation time for all but the lowest N_F . Furthermore, nearly all these states must be visited before converged relative free energies can be obtained, because the biased sampling removes any discrimination between these states in terms of their relative free energies in the physical ensemble. In other words, the FB-LEUS scheme with BF-type biasing potentials (and assuming that nonlocal effects are negligible) is not very different from a systematic or random scanning approach of the relevant conformational subspace. For example, scheme F_{30} applied to Ala_{10} produces about 230 000 unique peptide backbone conformations within 50 ns of simulation. The evaluation of as many conformations by systematic or random scanning within the same simulation time would permit a sampling time of 0.2 ps per conformation, which is very close to the effective visiting time of 0.34 ps for this simulation. The situation is different for the VD-type biasing potentials, where the relative free energies of the M different states at the fragment level remain unaltered compared with the physical system, and thus generally different (only the transition barriers are changed). As a result the corresponding M^{N_F} states at the real-system level

also have different free energies, and the biased sampling retains a statistical efficiency comparable to that of plain MD (and much higher than in the BF case).

Considering points B and D above, as well as the results of the present study, it appears that the use of the FB-LEUS scheme along with BF-type biasing potentials is not a viable approach to address high-dimensionality problems. On the other hand, the FB-LEUS scheme relying on VD-type biasing potentials successfully addresses points A, B, and D, while point C seems to be of little relevance in the context of the oligopeptide test systems considered here. Nevertheless, the results of the present study suggest that biasing potentials of this type do not represent a significant improvement over plain MD for poly-Ala and poly-Val oligopeptides.

The reason for this failure is probably in large part related to the choice of these test systems. Because the peptide linkage is relatively flexible, there is only little gain in conformational searching efficiency upon “digging” valleys between the corresponding free-energy basins. This moderate gain is in turn partly offset by a minimal but unavoidable loss of statistical efficiency upon opening a small irrelevant volume to sampling. In other words, for these systems, it appears nearly impossible to significantly improve over plain MD in terms of combined sampling efficiency.

Work is currently in progress to investigate the performance of the FB-LEUS approach in the context of oligosaccharide test systems. In view of the much more “rigid” nature of the glycosidic linkage,⁶² a significant sampling enhancement could be achieved in this case using VD-type biasing potentials. Finally, the development of an alternative approach, ball-and-stick LEUS (B&S-LEUS), that combines biasing potentials of low internal dimensionalities, minimal irrelevant volumes, and problem-adapted geometries (and is also generalizable to the calculation of “alchemical” free-energy differences) will be reported in a subsequent article.⁹⁰

Acknowledgment. Financial support from the Swiss National Science Foundation (Grant NF200021-121895) is gratefully acknowledged. X.D. also acknowledges support from the Spanish MICINN/FEDER (Grant BIO2007-62954).

Appendix A. LEUS Method with Truncated Polynomials

The local elevation umbrella sampling^{15,62} (LEUS) method consists of two steps: (i) a LE build-up (searching) phase that is used to construct an optimized memory-based biasing potential within a LE subspace of N_{LE} conformationally relevant degrees of freedom; (ii) an US sampling phase, where the (frozen) memory-based potential is used to generate a biased ensemble with extensive coverage of the US subspace defined by the same $N_{US} = N_{LE}$ degrees of freedom. During the LE build-up phase, the searching is carried out for a duration t_{LE} using (thermostatted and possibly barostatted) molecular dynamics (MD) based on the time-dependent potential energy function

$$\mathcal{U}_{LE}(\mathbf{r}, \mathcal{V}, t) = \mathcal{U}_{phys}(\mathbf{r}, \mathcal{V}) + \mathcal{U}_{bias}(\mathbf{Q}; \mathbf{M}(t)) \quad (\text{A.1})$$

During the subsequent US sampling phase, the sampling is carried out for a duration t_{US} using (thermostatted and

possibly barostatted) MD based on the time-independent potential energy function

$$\mathcal{U}_{US}(\mathbf{r}, \mathcal{V}) = \mathcal{U}_{phys}(\mathbf{r}, \mathcal{V}) + \mathcal{U}_{bias}(\mathbf{Q}; \mathbf{M}(t_{LE})) \quad (\text{A.2})$$

In eqs A.1 and A.2, \mathcal{U}_{phys} is the physical potential energy function of the system (force field), \mathcal{U}_{bias} is the memory-based biasing potential, $\mathcal{V} = \mathcal{V}(t)$ is the system volume at time t , $\mathbf{r} = \mathbf{r}(t)$ is the vector containing the Cartesian coordinates of all atoms in the system (configuration) at time t , $\mathbf{Q} = \mathbf{Q}(\mathbf{r})$ is the point representative of the current system configuration in the chosen LE subspace, and \mathbf{M} is the memory vector of the biasing potential. To entirely define the LEUS scheme, the above equations must be complemented by a representation of the biasing potential \mathcal{U}_{bias} in terms of its memory \mathbf{M} and by an updating scheme for this memory during the build-up phase, as discussed below.

The representation of the memory-based biasing potential relies on a discretization of the LE subspace (N_{LE} dimensions) by means of N_G grid points $\{\mathbf{Q}_n | n = 1, \dots, N_G\}$ defining the centers of nonoverlapping grid cells tiling this subspace. The memory \mathbf{M} consists of an N_G -dimensional integer vector accounting for the number of visits to a given grid cell (up to a time t during the build-up phase, or based on the total time t_{LE} of the build-up phase during the sampling phase). In its most general form, the biasing potential is then written

$$\mathcal{U}_{bias}(\mathbf{Q}; \mathbf{M}) = \sum_{n=1}^{N_G} k_{LE,n} M_n F_n(\mathbf{Q}) \quad (\text{A.3})$$

where F_n is a “unit” (i.e., normalized by the condition $F(0) = 1$) local repulsive function associated with grid point n and $k_{LE,n}$ is a corresponding force constant, while the updating scheme during the build-up phase is typically written

$$M_n(t + \Delta t) = M_n(t) + h_n(\mathbf{Q}) \quad (\text{A.4})$$

where Δt is the simulation time step and h_n evaluates to one for the (single) grid cell encompassing \mathbf{Q} and to zero otherwise. The grid-cell shapes and sizes, as well as the force constants $k_{LE,n}$ and functions F_n in eq A.3 can in principle be chosen differently for all grid points. However, unless such a high extent of flexibility is desired, it is common to assume a rectangular grid (spacing d_i along dimension i), a common force constant k_{LE} , and a unique functional form for F_n , the latter defined as a product of one-dimensional functions of the displacements relative to the grid-cell center along each dimension. In this case eq A.3 can be rewritten as

$$\mathcal{U}_{bias}(\mathbf{Q}; \mathbf{M}) = k_{LE} \sum_{n=1}^{N_G} M_n F(\mathbf{Q} - \mathbf{Q}_n; \mathbf{d}) \quad (\text{A.5})$$

with

$$F(\mathbf{Q}; \mathbf{d}) = \prod_{i=1}^{N_{LE}} f(Q_i; d_i) \quad (\text{A.6})$$

with the additional condition $f(0; d) = 1$ for any finite d .

The original LE method⁵⁶ relied on one-dimensional truncated Gaussian functions, that is,

$$f(x;d) = \exp\left[-\frac{x^2}{2\sigma^2}\right]H\left(1 - \frac{|x|}{d}\right) \quad (\text{A.7})$$

where H is the Heaviside step function (evaluating to one when its argument is positive, to zero otherwise), and σ is a Gaussian width ($\sigma = d$ was used in the original article⁵⁶). The use of eq A.7 is not recommended in practice because it leads to a biasing potential that is (i) generally not continuous and (ii) nonperiodic along possible periodic coordinates (e.g., angles).

The original LEUS method¹⁵ suggested the use of one-dimensional minimum-image Gaussian functions instead, that is,

$$f(x) = \exp\left[-\frac{\text{MI}(x)^2}{2\sigma^2}\right] \quad (\text{A.8})$$

where MI is the minimum-image function, selecting values within the period centered at zero (i.e., for which $|x|$ is minimal) for periodic coordinates, and σ is a Gaussian width ($\sigma = d$ was used in the original article¹⁵). Note that k_{LE} in eq A.5 is related to c_{LE} in ref 15 (see eqs 1 and 2 therein) as $k_{\text{LE}} = (2\pi\sigma^2)^{N_{\text{LE}}/2c_{\text{LE}}}$ (k_{LE} is a more convenient parameter, with units of energy). The use of eq A.8 represents an improvement, leading to a biasing potential that is (i) continuous and differentiable and (ii) periodic along possible periodic coordinates. However, because the Gaussian function is infinite-ranged, the formal range of the biasing potential contribution associated with a given grid point has been extended to the entire LE subspace (even if the corresponding effective range is much shorter when σ is chosen close to d). This leads to an increased computational cost or to the requirement of introducing a suitable cutoff distance (similar to eq A.7 but involving longer distances than d). Note that an alternative solution involves the multiplication of the Gaussian function by a polynomial that switches it to zero at finite range.⁴⁴

Clearly, neither eq A.7 nor eq A.8 is entirely satisfactory. In addition to the shortcomings mentioned above, grid-based Gaussian functions do not form an appropriate basis set for the representation of a constant function. This means that in addition to “flattening” the free-energy hypersurface in the relevant conformational subspace, the corresponding biasing potential will introduce spurious oscillations on a length scale equal to the grid spacing (these oscillations are clearly visible in Figure 2 of ref 56). The magnitude of these oscillations will increase with the extent of build-up and may ultimately result in a strong artificial bias of the system toward grid-cell boundaries, and the appearance of very high artificial forces preventing an accurate integration of the equations of motion. Finally, the calculation of a Gaussian-based biasing potential involves the evaluation of (computationally expensive) exponential functions. In the present study, truncated polynomial functions (similar to the assignment functions described in ref 91) are used instead, as detailed below.

The new formulation of the biasing potential employed in the present study relies on a local function of the form

$$f(x;d) = \left[1 - 3\left(\frac{\text{MI}(x)}{d}\right)^2 + 2\left(\frac{|\text{MI}(x)|}{d}\right)^3\right]H\left(1 - \frac{|\text{MI}(x)|}{d}\right) \quad (\text{A.9})$$

This local function satisfies the following desirable properties:

1. It is finite ranged, that is, $f(x) = 0$ for $|x| \geq d$, allowing for an evaluation of the biasing potential based on a sum involving a limited number of grid points ($2^{N_{\text{LE}}}$).
2. It is even, that is, $f(-x) = f(x)$, so that it does not induce any directional bias.
3. It is monotonic below and above $x = 0$, so that it does not induce artificial minima.
4. It is continuously differentiable (continuous first derivative), including at $x = \pm d$, leading to a continuously differentiable biasing potential.
5. It is periodic along periodic degrees of freedom, that is, $f(x + np) = f(x)$ where p is the period and n is a positive or negative integer (and assuming $p \geq d$).
6. It defines an appropriate grid-based basis set for the exact representation of the constant function, which follows from the property $f(x) + f(x - d) = 1$ for $0 < x < d$.
7. It evaluates to one at $x = 0$, that is, $f(0) = 1$, allowing for a direct interpretation of k_{LE} as the magnitude of the biasing energy at a grid point.
8. It is computationally inexpensive, since no exponential function is involved.

It is easily verified that if these one-dimensional functions define an appropriate grid-based basis set for the exact representation of the constant function, the same property holds for the function F of eq A.6 in the multidimensional LE subspace.

An alternative basis function satisfying properties 1–8 above and characterized in addition by a continuous second derivative would be

$$\tilde{f}(x;d) = \left[1 - 10\left(\frac{|\text{MI}(x)|}{d}\right)^3 + 15\left(\frac{\text{MI}(x)}{d}\right)^4 - 6\left(\frac{|\text{MI}(x)|}{d}\right)^5\right]H\left(1 - \frac{|\text{MI}(x)|}{d}\right) \quad (\text{A.10})$$

This alternative function could be employed in situations where integration errors caused by second-order discontinuities are an issue.⁹² Finally, it should be noted that an alternative approach involves the use of an interpolation scheme to obtain the bias energy between grid points.⁴⁸

Appendix B. Grid-Based BF and VD Biasing Potentials

In the FB-LEUS scheme, the build-up phase of duration t_{LE} is performed at the fragment level (LE subspace of dimension N_{LE}) according to eqs A.1 and A.4, using a memory \mathbf{M} of dimension N_{G} , resulting in an optimized biasing potential $\mathcal{Z}_{\text{bias}}(\mathbf{Q}; \mathbf{M}(t_{\text{LE}}))$. However, the sampling phase of duration t_{US} is performed at the system level (US subspace of dimension $N_{\text{US}} = N_{\text{F}}N_{\text{LE}}$ where N_{F} is the number of fragments in the system), replacing eq A.2 by

$$\mathcal{U}_{\text{US}}(\mathbf{r}, \mathcal{V}, t) = \mathcal{U}_{\text{phys}}(\mathbf{r}, \mathcal{V}) + \sum_{m=1}^{N_F} \mathcal{U}'_{\text{bias}}(\mathbf{Q}^{(m)}; h) \quad (\text{B.1})$$

Here, $\mathbf{Q}^{(m)} = \mathbf{Q}^{(m)}(\mathbf{r})$ is the N_{LE} -dimensional point representative of fragment m in the LE subspace and $\mathcal{U}'_{\text{bias}}$ is constructed according to the BF ($\mathcal{U}'_{\text{bias}}^{\text{BF}}$) or VD ($\mathcal{U}'_{\text{bias}}^{\text{VD}}$) procedure based on the given height parameter h (section 2.2). This construction requires the evaluation of the free energy $G(\mathbf{Q})$ from a sampling phase of duration t'_{US} at the fragment level, and stored as a N_{LE} -dimensional grid-based vector \mathbf{G} , that is,

$$G_n = -\beta^{-1} \ln \langle h_n(\mathbf{Q}) \exp[\beta \mathcal{U}'_{\text{bias}}(\mathbf{Q}; \mathbf{M}(t_{\text{LE}}))] \rangle_{t'_{\text{US}}} + C \quad (\text{B.2})$$

where C is chosen so that the lowest G_n value is zero (all others being positive). The procedure used to construct $\mathcal{U}'_{\text{bias}}^{\text{BF}}$ and $\mathcal{U}'_{\text{bias}}^{\text{VD}}$ based on $G(\mathbf{Q})$ is described qualitatively by eqs 1–5. A more precise (grid-based) description is provided below.

The grid-based expression corresponding to eq 1 is (in analogy to eq A.5)

$$\mathcal{U}'_{\text{bias}}^{\text{BF}}(\{\phi, \psi\}; h) = \sum_{n=1}^{N_G} M_n^{\text{BF}}(h) F_n(\{\phi - \phi_n, \psi - \psi_n\}; \{d_\phi, d_\psi\}) \quad (\text{B.3})$$

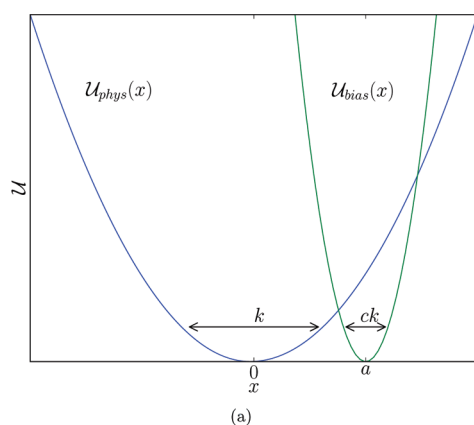
where ϕ_n and ψ_n are the ϕ and ψ values associated with grid point n , d_ϕ and d_ψ are the grid spacings along ϕ and ψ , and

$$M_n^{\text{BF}}(h) = \begin{cases} h - G_n & \text{if } G_n \leq h \\ 0 & \text{otherwise} \end{cases} \quad (\text{B.4})$$

The grid-based expression corresponding to eq 2 is (in analogy to eq A.5)

$$\mathcal{U}'_{\text{bias}}^{\text{VD}}(\{\phi, \psi\}; h) = \sum_{n=1}^{N_G} M_n^{\text{VD}}(h) F_n(\{\phi - \phi_n, \psi - \psi_n\}; \{d_\phi, d_\psi\}) \quad (\text{B.5})$$

with



$$M_n^{\text{VD}}(h) = \begin{cases} h - \Delta \tilde{G}_n & \text{if } \tilde{G}_n \leq h \\ 0 & \text{otherwise} \end{cases} \quad (\text{B.6})$$

where

$$\Delta \tilde{G}_n = \text{MAX}[\Delta G_{i(n,k)} - fj(n,k), k = 1, \dots, 8] \quad (\text{B.7})$$

In the above equations, MAX returns the maximum value of a set, $i(n, k)$ is the grid point defined by the projection of grid point n onto line k , $j(n, k)$ is the number of grid points separating n from $i(n, k)$ and f is a force threshold in units of energy per grid spacing. For a grid point belonging to a line k , $i(n, k) = n$ and $j(n, k) = 0$, so that eqs B.5–B.7 are equivalent to eq 2. For a grid point not belonging to a line k , these equations ensure that the force component transverse to a line will never exceed f . In practice, this modification will only affect lines parallel to a given line k and distant by one (or a few) grid spacing units, and predominantly in the high free energy region of the line. In the present work $d^{-1}f$ was set to 1 kJ mol⁻¹ deg⁻¹. Note that, as was the case for $\mathcal{U}'_{\text{bias}}$ in Appendix A, the biasing potentials $\mathcal{U}'_{\text{bias}}^{\text{BF}}$ and $\mathcal{U}'_{\text{bias}}^{\text{VD}}$ are weighted sums of continuously differentiable functions and are thus themselves continuously differentiable (despite the discontinuously differentiable truncation of the weights in eqs B.4 and B.6).

Appendix C. Significance of the Statistical Efficiency Factor

The statistical efficiency factor F introduced in eqs 8 and 9 characterizes the mutual relationship between the physical potential energy $\mathcal{U}_{\text{phys}}(\mathbf{r})$ and the biasing potential $\mathcal{U}_{\text{bias}}(\mathbf{r})$. This is most easily seen by considering the infinite-sampling limit of these equations. The configurational probability distributions within the physical and biased ensembles can be written

$$\rho_{\text{phys}}(\mathbf{r}) = Z_{\text{phys}}^{-1} \exp[-\beta \mathcal{U}_{\text{phys}}(\mathbf{r})] \quad (\text{C.1})$$

with

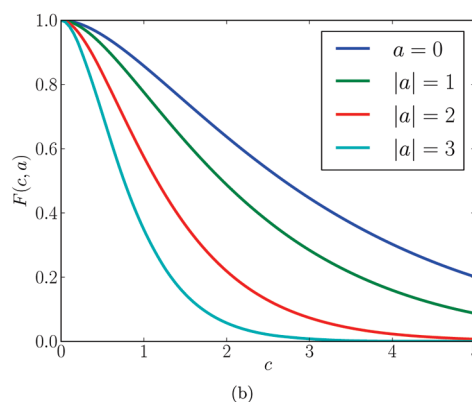


Figure 11. Illustration of the effect of the biasing potential on the statistical efficiency considering a simple one-dimensional harmonic situation: (a) physical potential energy $\mathcal{U}_{\text{phys}}(x)$, harmonic with force constant k , and biasing potential energy $\mathcal{U}_{\text{bias}}(x)$, harmonic with force constant ck and coordinate offset a (eqs C.10 and C.11); (b) statistical efficiency $F(c, a)$ as a function of c for different values of a (eq C.12).

$$Z_{\text{phys}} = \int d\mathbf{r} \exp[-\beta \mathcal{U}_{\text{phys}}(\mathbf{r})] \quad (\text{C.2})$$

and

$$\rho_{\text{biased}}(\mathbf{r}) = Z_{\text{biased}}^{-1} \exp\{-\beta[\mathcal{U}_{\text{phys}}(\mathbf{r}) + \mathcal{U}_{\text{bias}}(\mathbf{r})]\} \quad (\text{C.3})$$

with

$$Z_{\text{biased}} = \int d\mathbf{r} \exp\{-\beta[\mathcal{U}_{\text{phys}}(\mathbf{r}) + \mathcal{U}_{\text{bias}}(\mathbf{r})]\} \quad (\text{C.4})$$

Given a trajectory of N_f frames and in the limit $N_f \rightarrow \infty$, the density of frames in the biased ensemble is given by $N_f \rho_{\text{biased}}(\mathbf{r})$. Equations 8 and 9 can thus be rewritten in this limit as

$$F_n = N_f^{-1} \exp[-N_f \int d\mathbf{r} \rho_{\text{biased}}(\mathbf{r}) p(\mathbf{r}) \ln p(\mathbf{r})] \quad (\text{C.5})$$

with

$$p(\mathbf{r}) = \frac{N_f \int d\mathbf{r} \rho_{\text{biased}}(\mathbf{r}) \exp[\beta \mathcal{U}_{\text{bias}}(\mathbf{r})]^{-1} \exp[\beta \mathcal{U}_{\text{bias}}(\mathbf{r})]}{Z_{\text{biased}}} \exp[\beta \mathcal{U}_{\text{bias}}(\mathbf{r})] \quad (\text{C.6})$$

Using eqs C.1, C.4, and C.6, one can easily rearrange eq C.5 to

$$F = \frac{Z_{\text{phys}}}{Z_{\text{biased}}} \exp[-Z_{\text{phys}}^{-1} \int d\mathbf{r} \beta \mathcal{U}_{\text{bias}}(\mathbf{r}) \exp[-\beta \mathcal{U}_{\text{phys}}(\mathbf{r})]] \quad (\text{C.7})$$

or, in terms of ensemble averages $\langle \dots \rangle$ over the physical ensemble,

$$F = \frac{\exp[-\langle \beta \mathcal{U}_{\text{bias}}(\mathbf{r}) \rangle]}{\langle \exp[-\beta \mathcal{U}_{\text{bias}}(\mathbf{r})] \rangle} \quad (\text{C.8})$$

By applying the inequality of arithmetic and geometric means,⁹³ which states that

$$N^{-1} \sum_{i=1}^N x_i \geq \left(\prod_{i=1}^N x_i \right)^{1/N} \quad (\text{C.9})$$

when all $x_i > 0$ (the equality holding only when all x_i are identical), it is easily seen that the quantity F is always positive and reaches a maximum value of one for $\mathcal{U}_{\text{bias}}(\mathbf{r}) = cst$. This observation is important because it shows that the application of a biasing potential always decreases the statistical efficiency compared with a sampling relying on $\mathcal{U}_{\text{phys}}(\mathbf{r})$ only.

This behavior is illustrated qualitatively in Figure 11, considering a simple one-dimensional harmonic situation with

$$\mathcal{U}_{\text{phys}}(x) = \frac{1}{2} kx^2 \quad (\text{C.10})$$

and

$$\mathcal{U}_{\text{bias}}(x, c, a) = \frac{1}{2} ck(x - a)^2 \quad (\text{C.11})$$

In this case, eq C.8 can be evaluated analytically, leading to

$$F(c, a) = (c + 1)^{1/2} \exp\left[-\frac{c(c + 1 + \beta cka^2)}{2(c + 1)}\right] \quad (\text{C.12})$$

In the absence of biasing potential ($c = 0$), F evaluates to one as expected. However, as the value of c is increased (narrowing of the biasing potential), the statistical efficiency monotonically decreases toward a limiting value of zero. The decrease is more rapid when $|a|$ is large (decentering of the biasing potential), because the narrowing of the biasing potential focuses the sampling on high-energy regions in terms of the physical potential energy, but is also observed for $a = 0$ (centered biasing potential).

References

- (1) Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids*; Oxford University Press: New York, 1987.
- (2) van Gunsteren, W. F.; Berendsen, H. J. C. *Angew. Chem., Int. Ed.* **1990**, *29*, 992–1023.
- (3) van Gunsteren, W. F.; Bakowies, D.; Baron, R.; Chandrasekhar, I.; Christen, M.; Daura, X.; Gee, P.; Geerke, D. P.; Glättli, A.; Hünenberger, P. H.; Kastenholz, M. A.; Oostenbrink, C.; Schenk, M.; Trzesniak, D.; van der Vegt, N. F. A.; Yu, H. B. *Angew. Chem., Int. Ed.* **2006**, *45*, 4064–4092.
- (4) Berendsen, H. J. C. *Simulating the Physical World*; Cambridge University Press: Cambridge, U.K., 2007.
- (5) Rick, S. W.; Stuart, S. J. *Rev. Comput. Chem.* **2002**, *18*, 89–146.
- (6) Yu, H.; van Gunsteren, W. F. *Comput. Phys. Commun.* **2005**, *172*, 69–85.
- (7) Stern, H. A.; Berne, B. J. *J. Chem. Phys.* **2001**, *115*, 7622–7628.
- (8) Geerke, D. P.; Luber, S.; Marti, K. H.; van Gunsteren, W. F. *J. Comput. Chem.* **2008**, *30*, 514–523.
- (9) Hünenberger, P. H.; van Gunsteren, W. F. In *Computer Simulation of Biomolecular Systems, Theoretical and Experimental Applications*; van Gunsteren, W. F., Weiner, P. K., Wilkinson, A. J., Eds.; Kluwer/Escom Science Publishers: Dordrecht, The Netherlands, 1997; pp 3–82.
- (10) Kastenholz, M.; Hünenberger, P. H. *J. Phys. Chem. B* **2004**, *108*, 774–788.
- (11) Reif, M. M.; Kräutler, V.; Kastenholz, M. A.; Daura, X.; Hünenberger, P. H. *J. Phys. Chem. B* **2009**, *113*, 3112–3128.
- (12) van Gunsteren, W. F.; Huber, T.; Torda, A. E. *AIP Conf. Proc.* **1995**, *330*, 253–268.
- (13) Berne, B. J.; Straub, J. E. *Curr. Opin. Struct. Biol.* **1997**, *7*, 181–189.
- (14) Christen, M.; van Gunsteren, W. F. *J. Comput. Chem.* **2008**, *29*, 157–166.
- (15) Hansen, H. S.; Hünenberger, P. H. *J. Comput. Chem.* **2010**, *31*, 1–23.
- (16) Daura, X.; van Gunsteren, W. F.; Mark, A. E. *Proteins: Struct., Funct., Genet.* **1999**, *34*, 269–280.
- (17) Gnanakaran, S.; Nymeyer, H.; Portman, J.; Sanbonmatsu, K. Y.; Garcia, A. E. *Curr. Opin. Struct. Biol.* **2003**, *13*, 168–174.

- (18) Snow, C. D.; Sorin, E. J.; Rhee, Y. M.; Pande, V. S. *Annu. Rev. Biophys. Biomol. Struct.* **2005**, *34*, 43–69.
- (19) Eaton, W. A.; Muñoz, V.; Thompson, P. A.; Henry, E. R.; Hofrichter, J. *Acc. Chem. Res.* **1998**, *31*, 745–753.
- (20) Snow, C. D.; Nguyen, H.; Pande, V. S.; Gruebele, M. *Nature* **2002**, *420*, 102–106.
- (21) Williams, S.; Causgrove, T. P.; Gilmanshin, R.; Fang, K. S.; Callender, R. H.; Woodruff, W. H.; Dyer, R. B. *Biochemistry* **1996**, *35*, 691–697.
- (22) Lapidus, L. J.; Eaton, W. A.; Hofrichter, J. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 7220–7225.
- (23) Muñoz, V.; Thompson, P. A.; Hofrichter, J.; Eaton, W. A. *Nature* **1997**, *390*, 196–199.
- (24) Qiu, L.; Pabit, S. A.; Roitberg, A. E.; Hagen, S. J. *J. Am. Chem. Soc.* **2002**, *124*, 12952–12953.
- (25) Hünenberger, P. H.; Granwehr, J. K.; Aebischer, J.-N.; Ghoneim, N.; Haselbach, E.; van Gunsteren, W. F. *J. Am. Chem. Soc.* **1997**, *119*, 7533–7544.
- (26) Daura, X.; Jaun, B.; Seebach, D.; van Gunsteren, W. F.; Mark, A. E. *J. Mol. Biol.* **1998**, *280*, 925–932.
- (27) Torrie, G. M.; Valleau, J. P. *J. Comput. Phys.* **1977**, *23*, 187–199.
- (28) Valleau, J. P.; Torrie, G. M. In *Modern Theoretical Chemistry*; Berne, B. J., Ed.; Plenum Press: New York, 1977; Vol. 16, pp 9–194.
- (29) Beutler, T. C.; van Gunsteren, W. F. *J. Chem. Phys.* **1994**, *100*, 1492–1497.
- (30) Kumar, S.; Rosenberg, J. M.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A. *J. Comput. Chem.* **1995**, *16*, 1339–1350.
- (31) Heinz, T. N.; van Gunsteren, W. F.; Hünenberger, P. H. *J. Chem. Phys.* **2001**, *115*, 1125–1136.
- (32) Piccinini, E.; Ceccarelli, M.; Affinito, F.; Brunetti, R.; Jacoboni, C. *J. Chem. Theory Comput.* **2008**, *4*, 173–183.
- (33) Ferrenberg, A. M.; Swendsen, R. H. *Phys. Rev. Lett.* **1989**, *12*, 1195–1198.
- (34) Kumar, S.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A.; Rosenberg, J. M. *J. Comput. Chem.* **1992**, *13*, 1011–1021.
- (35) Bartels, C.; Karplus, M. *J. Comput. Chem.* **1997**, *18*, 1450–1462.
- (36) Chodera, J. D.; Swope, W. C.; Pitera, J. W.; Seok, C.; Dill, K. A. *J. Chem. Theory Comput.* **2007**, *3*, 26–41.
- (37) Kästner, J.; Thiel, W. *J. Chem. Phys.* **2005**, *123*, 144104.
- (38) Kästner, J.; Thiel, W. *J. Chem. Phys.* **2006**, *124*, 234106.
- (39) Paine, G. H.; Scheraga, H. A. *Biopolymers* **1985**, *24*, 1391–1436.
- (40) Mezei, M. *J. Comput. Phys.* **1987**, *68*, 237–248.
- (41) Hooft, R. W. W.; van Eijck, B. P.; Kroon, J. *J. Chem. Phys.* **1992**, *97*, 6690–6694.
- (42) Friedman, R. A.; Mezei, M. *J. Chem. Phys.* **1995**, *102*, 419–426.
- (43) Wang, J.; Gu, Y.; Liu, H. *J. Chem. Phys.* **2006**, *125*, 094907.
- (44) Babin, V.; Roland, C.; Darden, T. A.; Sagui, C. *J. Chem. Phys.* **2006**, *125*, 204909.
- (45) Marsili, S.; Barducci, A.; Chelli, R.; Procacci, P.; Schettino, V. *J. Phys. Chem. B* **2006**, *110*, 14011–14013.
- (46) Lelièvre, T.; Rousset, M.; Stoltz, J. *Chem. Phys.* **2007**, *126*, 134111.
- (47) van der Vaart, A.; Karplus, M. *J. Chem. Phys.* **2007**, *126*, 164106.
- (48) Babin, V.; Roland, C.; Sagui, C. *J. Chem. Phys.* **2008**, *128*, 134101.
- (49) Barnett, C. B.; Naidoo, K. J. *Mol. Phys.* **2009**, *107*, 1243–1250.
- (50) Laio, A.; Parrinello, M. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 12562–12566.
- (51) Laio, A.; Rodriguez-Fortea, A.; Gervasio, F. L.; Ceccarelli, M.; Parrinello, M. *J. Phys. Chem. B* **2005**, *109*, 6714–6721.
- (52) Darve, E.; Rodriguez-Gomez, D.; Pohorille, A. *J. Chem. Phys.* **2008**, *128*, 144120.
- (53) Crippen, G. M.; Scheraga, H. A. *Chemistry* **1969**, *64*, 42–49.
- (54) Levy, A. V.; Montalvo, A. *SIAM J. Sci. Stat. Comput.* **1985**, *6*, 15–29.
- (55) Glover, F. *ORSA J. Comput.* **1989**, *1*, 190–206.
- (56) Huber, T.; Torda, A. E.; van Gunsteren, W. F. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 695–708.
- (57) Grubmüller, H. *Phys. Rev. E* **1995**, *52*, 2893–2906.
- (58) Engkvist, O.; Karlström, G. *Chem. Phys.* **1996**, *213*, 63–76.
- (59) Fukunishi, Y.; Mikami, Y.; Nakamura, H. *J. Phys. Chem. B* **2003**, *107*, 13201–13210.
- (60) van Gunsteren, W. F.; Billeter, S. R.; Eising, A. A.; Hünenberger, P. H.; Krüger, P.; Mark, A. E.; Scott, W. R. P.; Tironi, I. G. *Biomolecular Simulation: The GROMOS96 Manual and User Guide*; Verlag der Fachvereine: Zürich, Switzerland, 1996.
- (61) Scott, W. R. P.; Hünenberger, P. H.; Tironi, I. G.; Mark, A. E.; Billeter, S. R.; Fennen, J.; Torda, A. E.; Huber, T.; Krüger, P.; van Gunsteren, W. F. *J. Phys. Chem. A* **1999**, *103*, 3596–3607.
- (62) Perić-Hassler, L.; Hansen, H. S.; Baron, R.; Hünenberger, P. H. *Carbohydr. Res.* **2010**, *345*, 1781–1801.
- (63) Leitgeb, M.; Schröder, C.; Boresch, S. *J. Chem. Phys.* **2005**, *122*, 084109.
- (64) Darve, E.; Pohorille, A. *J. Chem. Phys.* **2001**, *115*, 9169–9183.
- (65) Darve, E.; Wilson, M. A.; Pohorille, A. *Mol. Simul.* **2002**, *28*, 113–144.
- (66) Naidoo, K. J.; Brady, J. W. *J. Am. Chem. Soc.* **1999**, *121*, 2244–2252.
- (67) Kuttel, M. M.; Naidoo, K. J. *J. Phys. Chem. B* **2005**, *109*, 7468–7474.
- (68) Strümpfer, J.; Naidoo, K. J. *J. Comput. Chem.* **2010**, *31*, 308–316.
- (69) Ensing, B.; Laio, A.; Parrinello, M.; Klein, M. L. *J. Phys. Chem. B* **2005**, *109*, 6676–6687.
- (70) Li, H.; Min, D.; Liu, Y.; Yang, W. *J. Chem. Phys.* **2007**, *127*, 094101.
- (71) Hansen, H. S.; Hünenberger, P. H. *J. Comput. Chem.* 2010, submitted for publication.
- (72) Kannan, S.; Zacharias, M. *Proteins: Struct., Funct., Bioinf.* **2007**, *66*, 697–706.
- (73) Xu, C.; Wang, J.; Liu, H. *J. Chem. Theory Comput.* **2008**, *4*, 1348–1359.

- (74) Christen, M.; Hünenberger, P. H.; Bakowies, D.; Baron, R.; Bürgi, R.; Geerke, D. P.; Heinz, T. N.; Kastenholz, M. A.; Kräutler, V.; Oostenbrink, C.; Peter, C.; Trzesniak, D.; van Gunsteren, W. F. *J. Comput. Chem.* **2005**, *26*, 1719–1751.
- (75) Oostenbrink, C.; Villa, A.; Mark, A. E.; van Gunsteren, W. F. *J. Comput. Chem.* **2004**, *25*, 1656–1676.
- (76) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Hermans, J. In *Intermolecular Forces*; Pullman, B., Eds.; Reidel: Dordrecht, The Netherlands, 1981; Vol. 33, pp 1–342.
- (77) Feynman, R. P.; Leighton, R. B.; Sands, M. *The Feynman Lectures on Physics*; Addison-Wesley: Boston, MA, 1963.
- (78) Hockney, R. W. *Methods Comput. Phys.* **1970**, *9*, 136–211.
- (79) Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *23*, 327–341.
- (80) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Di Nola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684–3690.
- (81) Berendsen, H. J. C.; van Gunsteren, W. F.; Zwinderman, H. R. J.; Geurtsen, R. G. *Ann. N.Y. Acad. Sci.* **1986**, *482*, 269–285.
- (82) Tironi, I. G.; Sperb, R.; Smith, P. E.; van Gunsteren, W. F. *J. Chem. Phys.* **1995**, *102*, 5451–5459.
- (83) Weber, W.; Hünenberger, P. H.; McCammon, J. A. *J. Phys. Chem. B* **2000**, *104*, 3668–3675.
- (84) Wu, D.; Kofke, D. A. *J. Chem. Phys.* **2005**, *123*, 054103.
- (85) Wu, D.; Kofke, D. A. *J. Chem. Phys.* **2005**, *123*, 084109.
- (86) Shen, T.; Hamelberg, D. *J. Chem. Phys.* **2008**, *129*, 034103.
- (87) Shell, M. S. *J. Chem. Phys.* **2008**, *129*, 144108.
- (88) Zwanzig, R.; Szabo, A.; Bagchi, B. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 20–22.
- (89) Kabsch, W.; Sander, C. *Biopolymers* **1983**, *22*, 2577–2637.
- (90) Hansen, H. S.; Hünenberger, P. H. *J. Chem. Theory Comput.*, DOI: 10.1021/ct1003065.
- (91) Hünenberger, P. H. In *Simulation and Theory of Electrostatic Interactions in Solution: Computational Chemistry, Biophysics, and Aqueous Solution*; Hummer, G., Pratt, L. R., Eds.; American Institute of Physics: New York, 1999; Vol. 1, pp 7–83.
- (92) Hamelberg, D.; Mongan, J.; McCammon, J. A. *J. Chem. Phys.* **2004**, *120*, 11919–11929.
- (93) Jensen, J. L. W. V. *Acta Math.* **1906**, *30*, 175–193.
CT1003059

Ball-and-Stick Local Elevation Umbrella Sampling: Molecular Simulations Involving Enhanced Sampling within Conformational or Alchemical Subspaces of Low Internal Dimensionalities, Minimal Irrelevant Volumes, and Problem-Adapted Geometries

Halvor S. Hansen and Philippe H. Hünenberger*

Laboratorium für Physikalische Chemie, ETH Zürich, CH-8093 Zürich, Switzerland

Received June 5, 2010

Abstract: A new method, ball-and-stick local elevation umbrella sampling (B&S-LEUS), is proposed to enhance the sampling in computer simulations of (bio)molecular systems. It enables the calculation of conformational free-energy differences between states (or alchemical free-energy differences between molecules), even in situations where the definition of these states relies on a conformational subspace involving more than a few degrees of freedom. The B&S-LEUS method consists of the following steps: (A) choice of a reduced conformational subspace; (B) representation of the relevant states by means of spheres (“balls”), each associated with a biasing potential involving a one-dimensional radial memory-based term and a radial confinement term; (C) definition of a set of lines (“sticks”) connecting these spheres, each associated with a biasing potential involving a one-dimensional longitudinal memory-based term and a transverse confinement term; (D) unification of the biasing potentials corresponding to the union of all of the spheres and lines (active subspace) into a single biasing potential according to the enveloping distribution sampling (EDS) scheme; (E) build-up of the memory using the local elevation (LE) procedure, leading to a biasing potential enabling a nearly uniform sampling (radially within the spheres, longitudinally within the lines) of the active subspace; (F) generation of a biased ensemble of configurations using this preoptimized biasing potential, following an umbrella sampling (US) approach; and (G) calculation of the relative free energies of the states via reweighting and state assignment. The main characteristics of this approach are: (i) a low internal dimensionality, that is, the memory only involves one-dimensional grids (acceptable memory requirements); (ii) a minimal irrelevant volume, that is, the conformational volume opened to sampling includes a minimal fraction of irrelevant regions in terms of the free energy of the physical system or of user-specified metastable states (acceptable build-up duration requirements, high statistical efficiency); and (iii) a problem-adapted geometry (a priori specification of the conformational regions considered as relevant or irrelevant). In particular, the use of lines to connect the spheres ensures both a minimal irrelevant volume and a sufficient number of transitions between the states. As an illustration, the B&S-LEUS method is applied here to three test systems: (i) a solvated (blocked) alanine mono-peptide (two-dimensional conformational subspace), used as a toy system to illustrate the versatility of the method in promoting the sampling of arbitrary regions of the Ramachandran map; (ii) a solvated polyalanine decapeptide (nine-dimensional conformational subspace), to evaluate the relative free energies of three different types of helices (π , α , and 3_{10}) based on a single simulation; and (iii) a solvated artificial hexopyranose, termed the “mother” of all D-hexopyranoses and constructed as a hybrid of all D-hexopyranose stereoisomers, where the method is applied (seven-dimensional mixed alchemical and conformational subspace) to calculate the relative free energies of the corresponding 32 isomers, anomers, and chair conformers, based on a single simulation.

1. Introduction

Classical atomistic simulations, and in particular molecular dynamics (MD), represent nowadays a powerful tool complementary to experiment for investigating the properties of molecular systems relevant in physics, chemistry, and biology.^{1–4} Their success in the context of condensed-phase systems results in particular from a favorable trade-off between the spatial and temporal resolutions of these models (on the order 0.1 nm and 1 fs) and their computational costs, permitting to reach system sizes and time scales relevant for many (bio)molecular applications (on the order of 10 nm and 100 ns). These scales are sufficient to enable in many cases: (i) an appropriate description of bulk-like solvation using discrete solvent molecules; (ii) a reliable calculation of thermodynamic properties via statistical mechanics; and (iii) a direct comparison of simulated properties with experimental data.

In practice, however, the results of atomistic simulations are still affected by four main sources of error, originating from: (i) the classical atomistic approximation;^{5–8} (ii) the approximate force-field representation of interatomic interactions;^{2–4,9} (iii) the presence of finite-size and surface effects;^{10,11} and (iv) the insufficient conformational (and, possibly, alchemical; see below) sampling.^{9,12–15} The reduction of the last type of errors can be viewed as a first-priority target in the improvement of simulation methodologies, because these errors are predominantly nonsystematic, while the three other types of errors are systematic. The present project is concerned with the design of a new sampling-enhancement scheme with the goal of reducing the corresponding errors in the context of free-energy calculations.

Free-energy calculations based on classical atomistic simulations^{2,3,16–23} can be classified into two main categories. On the one hand, the calculation of conformational free energies aims at evaluating the relative free energies of relevant conformational states of a given molecular system (as well as corresponding free-energy profiles or maps). Typical examples include the evaluation of the relative free energies of the bound and unbound states of a molecular complex (e.g., protein–ligand complex^{24,25}), or of conformational states of a macromolecule presenting different spectroscopic or functional properties (e.g., folded and unfolded states of a protein,²⁶ α -helical or β -sheet conformation of a peptide,²⁷ different double-helical forms of an oligonucleotide²⁸). On the other hand, the calculation of alchemical free energies aims at evaluating the relative free energies of different molecules (or, more precisely, molecular topologies) in a given environment (the corresponding profiles or maps are then unphysical and irrelevant). In general, the target quantity is actually in this case a difference between the relative free energies calculated considering two different environments, so as to characterize the environmental effect via a thermodynamic cycle. Typical examples include the evaluation of solvation free energies^{29–31} (difference between the free-energy change upon “creating” a molecule in solution and in the gas phase) or of relative

binding free energies^{32,33} (difference between the free-energy change upon “mutating” a molecule into another one in solution and within a molecular complex). Note that the “creation” or “mutation” must be performed at constant number of atoms, possibly requiring the introduction of “dummy” atoms (covalently linked mass sites free of nonbonded interactions). Finally, for completeness, one might mention a third category of free-energy calculations, involving thermodynamic free-energy changes (i.e., free-energy changes upon variation of a thermodynamic parameter such as temperature or pressure).

In the early days of free-energy calculations, the evaluation of relative free energies was typically restricted to two-states or two-molecules problems, that is, to the evaluation of a single free-energy difference.

The established methods for the calculation of a single conformational free-energy difference are direct counting³⁴ (DC) and umbrella sampling^{35,36} (US). The DC approach is only applicable in the (uncommonly) favorable situation where the dynamics spontaneously samples the two conformational states with a sufficient number of interconversion transitions. In this case, the free-energy difference can be calculated directly from the ratio of the numbers of sampled conformations assigned to either of the two states. The US approach is more generally applicable and relies on the use of a time-independent biasing potential that forces the sampling of the two states with a sufficient number of interconversion transitions. In this case, the free-energy difference can be calculated from the ratio of the reweighted numbers of conformations assigned to either of the two states, the reweighting acting as a correction for the effect of the biasing. In practice, the direct design of a biasing potential satisfying the required properties for two significantly differing conformational states is difficult, and one has to resort to³⁷ multiple-windows,^{38–41} adaptive,^{42–52} or memory-based^{15,53–56} approaches.

The established methods for the calculation of a single alchemical free-energy difference are thermodynamic integration^{57,58} (TI) and free-energy perturbation^{59,60} (FP). In both cases, a hybrid Hamiltonian is constructed by introduction of a coupling parameter λ , in such a way that $\lambda = 0$ corresponds to the initial molecule and $\lambda = 1$ to the final molecule. In the TI approach, the free-energy difference is obtained from the ensemble average of the Hamiltonian λ -derivative evaluated at a discrete set of successive λ -points, via numerical integration. In the FP approach, the free-energy difference is obtained on the basis of an ensemble average performed at a single λ -point, involving the relative Boltzmann weights of the Hamiltonians associated with the two molecules. In practice, the latter scheme becomes inaccurate in the case of two significantly differing molecules, and a multiple-windows¹⁸ approach must be used instead.

The above methods for the calculation of single free-energy differences are still widely used nowadays and, after the resolution of some important methodological issues in the 1990s (Hamiltonian lag,⁶¹ singularity upon atom creation and deletion,^{62,63} metric-tensor effects,^{64–66} contribution of constraints,^{66–68} contribution of restraints,^{24,25,69} standard-state corrections^{24,25}), do not present major difficulties. Their

* Corresponding author phone: +41 44 632 5503; fax: +41 44 632 1039; e-mail: phil@igc.phys.chem.ethz.ch.

main drawback is that they are restricted to two-state or two-molecule (i.e., pairwise) problems (low throughput) and typically require a significant amount of human time (multiple simulations, case-to-case adjustment of the staging, equilibration and sampling protocols, analysis of the data). Recently, however, the field of free-energy calculation has witnessed three important evolutions, which are described in turn below.

As a first recent development, research has turned to the new challenge of calculating multiple free-energy differences from a single simulation (or, possibly, from a set of simulations generated in an “automated” way). This challenge is particularly relevant in the context of alchemical changes, for example, for the calculation of the relative binding free energies of a large collection of possible ligands to the same receptor simultaneously (a “holy grail” for drug design^{70,71}). Initial attempts, relying on the application of the FP formula in an extrapolative way (one-step perturbation) based on a real molecule as reference state, were not very successful.^{72–74} Subsequent attempts, relying on the design of an unphysical reference state instead (e.g., molecule with soft-sites⁶³ aiming at encompassing configurations representative of all possible final states within the reference ensemble), were in specific cases more successful,^{32,75–77} but remained generally speaking moderately reliable. Arguably, the first practically useful approach of this kind is enveloping distribution sampling^{33,78–80} (EDS), where an unphysical reference-state Hamiltonian is constructed automatically, which presents an optimal overlap with (and sufficient interconversion transitions between) the different target molecules. Note that the EDS method has only been applied until now to alchemical free-energy calculations and remains to be generalized to conformational problems. One general lesson from this research is that viable approaches for the calculation of multiple free-energy differences based on a single reference simulation require the targets (states or molecules) to be known a priori, that is, purely extrapolative approaches are not very reliable in practice.

As a second recent development, efficient MD-based methods have been developed to address the conformational-searching problem,^{14,15} that is, the problem of scanning a potential energy hypersurface for low-energy configurations over the widest possible volume. Probably the most efficient types of MD-based searching methods available nowadays are those that rely on the progressive build-up of a memory-based penalty potential, preventing the continuous revisiting of previously discovered configurations. Many closely related variants of this approach can be found in the literature, including (chronologically) the deflation,⁸¹ tunneling,⁸² tabu search,⁸³ local elevation,⁸⁴ conformational flooding,⁸⁵ Engkvist–Karlström,⁸⁶ adaptive reaction coordinate force,⁵³ adaptive biasing force,⁵⁶ metadynamics,^{54,55} and filling potential⁸⁷ methods. The first practically useful implementation of this approach in the context of (bio)molecular systems with explicit solvation is probably the local elevation (LE) method of Huber, Torda, and van Gunsteren,⁸⁴ as implemented in the GROMOS96 program.^{88,89} In this method, the searching enhancement is applied along a subset of degrees of freedom of the system (LE-subspace), typically a limited

set of conformationally relevant dihedral angles, by means of a penalty potential defined as a sum of local (grid-based) repulsive functions, the magnitudes of which are made proportional to the number of previous visits to the corresponding conformation (grid cell). Because memory-based searching methods (such as the LE method) have a time-dependent Hamiltonian, they sample in principle no well-defined configurational probability distribution, that is, the resulting trajectories cannot be used for the evaluation of thermodynamic properties (including free energies) via statistical mechanics. However, in view of their very high searching power, there has been a long-standing interest in using their basic principle to design efficient conformational-sampling methods, that is, leading to trajectories suited for the evaluation of thermodynamic properties. This can be done by observing that at the end of a memory-based search, the penalty potential has approximately “flattened” the free-energy hypersurface in the considered subspace up to a certain threshold value above the lowest minimum discovered.⁸⁶ As a result, this final penalty potential represents an optimal biasing potential for a subsequent US simulation. Such a combination is at the heart of the local elevation umbrella sampling^{19,50} (LEUS) method (see below). Note that the LEUS scheme has only been applied until now to conformational free-energy calculations and remains to be generalized to alchemical problems.

Finally, as a third recent development, the so-called λ -dynamics approach^{91–96} has been proposed, in which the λ -variable of an alchemical change (possibly generalized to a λ -vector in the context of multiple changes) evolves dynamically in time along with the physical (atomic) degrees of freedom of the system (extended Lagrangian approach). Taken alone, λ -dynamics does not work very well in practice, because the sampling of the λ -space: (i) is heavily biased toward the regions corresponding to the most stable (lowest free energy) molecules; (ii) is generally hindered by the presence of free-energy barriers and local free-energy minima; and (iii) opens up a large volume of unphysical (thus irrelevant) alchemical space (in the context of multiple free-energy changes), thereby reducing the statistics relevant for the physical molecules. However, this approach has an important merit. It shows that any alchemical free-energy calculation can be reformulated as a pseudo-conformational calculation in an extended space including the λ -variables. In other words, methods that have been developed for the evaluation of conformational free-energy differences (e.g., DC, US and LEUS) can as well be applied to alchemical changes in the context of an extended-system dynamics. Note that a similar principle underlies the combination of LEUS with EDS for alchemical changes, as used previously⁹⁷ and in the present work (see below). The inverse mapping, that is, the application of methods that have been developed for the evaluation of alchemical free-energy changes (e.g., TI, FE, and EDS) to conformational changes, is also in principle possible, requiring the introduction of constraints along the relevant physical degrees of freedom of the system.^{14,28,68,98–100} However, because the constrained degrees of freedom are generally non-Cartesian coordinates, the resulting (projected) conformational subspace is typically non-Euclidean, which

raises a number of issues related to the Jacobian of the transformation and the possible occurrence of metric-tensor effects in MD simulations.^{64–66,98,101,102} These problems do not arise when extending a physical system to include an alchemical subspace, as long as the resulting extended space is chosen to be Euclidean. This is always possible by construction (provided that the physical space itself is Euclidean, i.e., in the absence of constrained internal coordinates) given the unphysical (thus irrelevant) nature of the geometry selected for this alchemical space.

Considering the above discussion, the LEUS method (including its generalization by combination with λ -dynamics or EDS) appears to represent a powerful scheme for the evaluation of both conformational and alchemical free-energy changes, including the determination of multiple changes from a single simulation.

As detailed in the original article,¹⁵ the LEUS scheme consists of two steps: (i) a LE build-up (searching) phase, that is used to progressively construct an optimized memory-based biasing potential within a LE-subspace of N_{LE} conformationally relevant degrees of freedom; and (ii) an US sampling phase, where this potential, now frozen, is used to generate a biased ensemble with extensive coverage of the US-subspace defined by the same $N_{US} = N_{LE}$ degrees of freedom. A successful build-up phase will produce a biasing potential that is approximately equal to the negative of the free-energy hypersurface within the considered subspace up to a certain free-energy level, so that a sufficiently long sampling phase will result in a nearly homogeneous coverage of the corresponding region. In addition, because the biasing potential in this second phase is time-independent, thermodynamic information relevant for the physical (unbiased) ensemble can be recovered from the simulated data by means of a simple reweighting procedure.^{15,35,36}

The LEUS scheme is a powerful sampling-enhancement technique in cases where the relevant conformational subspace is of low dimensionality.^{15,37,90} However, this scheme becomes inapplicable in its original form for systems where the dimension of this subspace exceeds a few degrees of freedom. If the relevant subspace is defined by $N_{LE} = N_{US} = N$ degrees of freedom, each degree of freedom being discretized by means of N_g grid points, and the biasing potential is expected to map out a fraction f of this subspace, the number of local functions required is $f(N_g)^N$. This number increases exponentially with N , so that the original LEUS approach rapidly becomes intractable, in terms of both memory and build-up duration requirements. A tentative solution to this problem,³⁷ fragment-based LEUS (FB-LEUS), relies on the preoptimization of fragment-based biasing potentials of low dimensionalities, followed by their simultaneous application to each of the corresponding fragments in a molecule, based on a similar principle as suggested in refs 103 and 104. This corresponds to a situation where $N_{US} = N_F N_{LE}$, where N_F is the number of fragments in the considered molecule. In principle, the resulting biasing potential should remain appropriate in situations where the free-energy function in the N_{US} -dimensional relevant subspace of the molecule can be approximated as a sum of N_{LE} -dimensional fragment-based contributions, thereby neglecting

the corresponding correlations. In this case, the fragment-based biasing potential will still lead to a nearly homogeneous coverage of this subspace, while the neglected correlations are reintroduced during the reweighting procedure. Application of the FB-LEUS scheme to solvated polyalanine and polyvaline oligopeptides using biasing potentials designed for the corresponding (blocked) mono-peptide fragments confirmed the above suggestion in the context of these systems.³⁷ However, it revealed another problem, namely that the enhancement of the searching power (volume of conformational space visited in a given amount of simulation time) is largely offset by a deterioration of the statistical efficiency (representativeness of the biased ensemble in terms of the conformational distribution appropriate for the physical ensemble). If the “flattening” of the free-energy hypersurface largely increases the rate at which new conformations are generated, it also includes a very large number of high free-energy conformations into the sampling. This does not represent a problem for low-dimensionality systems, where the simulation time scale is sufficient to afford the sampling of these extra configurations while maintaining reasonable statistics concerning the low free-energy regions. However, this is no longer the case for high-dimensionality problems, where the combinatorial crowding of the biased ensemble with irrelevant configurations leaves virtually no room for statistics concerning the relevant ones. This situation was tentatively remedied³⁷ by the introduction of alternative biasing potentials inducing the “digging” of valleys between the relevant conformational states at the fragment level, rather than the “flattening” of the corresponding conformational basins. While possibly interesting in the context of other polymers (e.g., oligosaccharides), these potentials did not result in a significant sampling enhancement (as compared to plain MD) for the considered oligopeptides, probably due to the already low conformational transition barriers at the level of the corresponding peptide linkages.

Despite being moderately successful per se, this study³⁷ suggested that an efficient memory-based biasing potential for high-dimensional problems should possess the following three characteristics: (i) it should be of low internal dimensionality (acceptable memory requirements); (ii) it should map out a minimal irrelevant volume (acceptable build-up duration requirements, high statistical efficiency); and (iii) it should possess a problem-adapted geometry (involving a priori specification of the conformational regions considered as relevant or irrelevant). The internal dimensionality refers to the dimensionality of the involved memory map, possibly being inscribed within a relevant conformational subspace of much higher dimensionality (e.g., one-dimensional curve within a multidimensional subspace). The irrelevant volume refers to the volume that is neither useful to the sampling of the relevant states nor strictly necessary to ensure a sufficient number of transitions between them. The above requirements can only be satisfied if the states considered to be relevant are defined prior to the design of the biasing potential, that is, this potential must possess a problem-adapted geometry. The above conditions form the basis of the new scheme proposed in the present work. This scheme is termed ball-

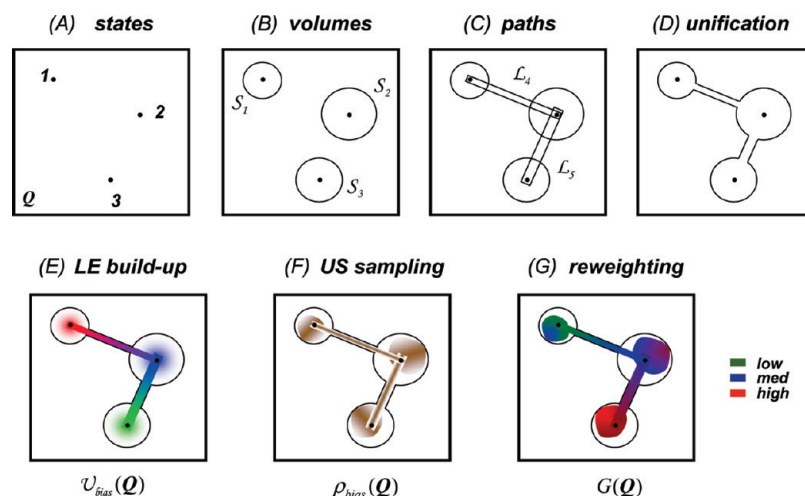


Figure 1. Schematic illustration of the successive steps involved in the B&S-LEUS approach. The reduced conformational space \mathbf{Q} is shown as a two-dimensional plane, and the calculation involves three relevant states (1, 2, and 3). The corresponding $K = 3$ conformational spheres are labeled \mathcal{S}_1 , \mathcal{S}_2 , and \mathcal{S}_3 . They are connected by $L = 2$ conformational lines, labeled \mathcal{L}_4 and \mathcal{L}_5 . In this illustration, the ranking of the states in terms of increasing free energy is $1 < 2 < 3$.

and-stick LEUS (B&S-LEUS), by reference to the assembly of sampling “balls” (encompassing the relevant states) and “sticks” (used to promote transitions between them) typically involved in the definition of the corresponding biasing potential.

More specifically, the B&S-LEUS approach aims at: (i) the extension of the LEUS scheme to the evaluation of the relative free energies of conformational states in high-dimensionality problems, that is, in cases where the plain LEUS¹⁵ or FB-LEUS³⁷ methods fail; (ii) its extension to the simultaneous determination of multiple conformational free-energy differences from the sampling phase of a single simulation; and (iii) taking a first step toward the generalization of the LEUS scheme to alchemical free-energy calculations, here in conjunction with the EDS scheme. This is achieved by simultaneously taking advantage of the explorative power of the LE searching procedure,⁸⁴ the focusing properties of restraining potentials, the Hamiltonian combination capability of the EDS method,⁷⁸ and the statistical correctness (time-independent Hamiltonian) of the US method.³⁵

In practice, the B&S-LEUS procedure relies on the following steps, illustrated schematically in Figure 1: (A) choice of a reduced conformational subspace permitting the definition of the relevant conformational states; (B) representation of the relevant conformational states by means of centered volumes within this subspace (e.g., conformational spheres or “balls”), each associated with a biasing potential involving a one-dimensional radial memory-based term and a radial confinement (half-harmonic restraining) term; (C) definition of a set of conformational paths connecting the centers of these volumes (e.g., conformational lines or “sticks”), each associated with a biasing potential involving a one-dimensional longitudinal memory-based term and a transverse confinement (flat-bottom half-harmonic restraining) term; (D) unification of the biasing potentials associated with all of the centered volumes and paths into a single biasing potential according to the EDS procedure; (E) optimization of the memory, leading to a biasing potential

enabling a nearly homogeneous sampling (radially within the centered volume, longitudinally within the paths) of the subvolume of the conformational subspace defined by the union of all centered volumes and paths (LE build-up phase); (F) generation of a biased ensemble of configurations using this preoptimized biasing potential (US sampling phase); and (G) calculation of the relative free energies of the states (reweighting and state-assignment procedure).

This Article is organized as follows. Section 2 describes in detail the successive steps (A–G) of the B&S-LEUS scheme. Section 3 introduces the three systems considered here to assess the performance of the scheme and provides the corresponding computational details. These three test systems are: (i) a solvated (blocked) alanine monopeptide (two-dimensional conformational subspace), used as a toy system to illustrate the versatility of the method in promoting the sampling of arbitrary regions of the Ramachandran map; (ii) a solvated (unblocked) polyalanine decapeptide (nine-dimensional conformational subspace), where the method is applied to evaluate the relative free energies of three different types of helices (π , α , and 3_{10}); (iii) a solvated artificial hexopyranose, termed the “mother” of all D-hexopyranoses and constructed as a hybrid of all D-hexopyranose stereoisomers, where the method is applied (seven-dimensional mixed alchemical and conformational subspace) to calculate the relative free energies of the corresponding 32 isomers, anomers, and chair conformers from a single simulation (8 stereoisomers, α - or β -anomer, 4C_1 or 1C_4 chair conformer). Section 4 reports the results of the B&S-LEUS simulations considering these three test systems. Finally, section 5 provides concluding remarks.

2. Method

This section describes the successive steps of the B&S-LEUS scheme, with reference to section 1 (points A–G) and Figure 1.

2.1. Choice of a Reduced Conformational Subspace.

The first step (A) in the B&S-LEUS approach relies on the choice of a reduced conformational subspace permitting the definition of the relevant conformational states. This is done by selecting a subset of N internal coordinates of the physical system, collectively noted by the vector $\mathbf{Q} = \{Q_n, n = 1, \dots, N\}$. These coordinates are assumed to be well-defined and differentiable functions of the vector \mathbf{r} encompassing the Cartesian coordinates of all particles in the system, that is, the vector function $\mathbf{Q} = \mathbf{Q}(\mathbf{r})$ must be defined for any \mathbf{r} and its derivative must be nonsingular.

Examples of possible internal coordinates include, for example, distances between atom pairs, angles between atom triples, dihedral angles between atom quadruples, root-mean-square atomic positional deviations from given reference structures, extended-system variables (e.g., λ -variables in λ -dynamics^{91,92,105}), or any (differentiable) mathematical combination of these. Note that periodic internal coordinates (e.g., angles) should not be “refolded” to a reference period, that is, their time evolution must be continuous. Furthermore, it is assumed that the definition of any internal coordinate (with a specified unit) is associated with the selection of a corresponding reference value σ_n (with the same unit), and that Q_n is defined by the unitless ratio of the two quantities. It is important to stress that the results of a B&S-LEUS simulation depend on a given choice of the σ_n factors, so that these factors must be clearly specified as an integral part of the definition of the conformational subspace.

2.2. Representation of the Relevant Conformational States. The second step (B) in the B&S-LEUS approach relies on the representation of the relevant conformational states by means of K centered volumes within the reduced conformational subspace. Only the simplest possible type of centered volume will be considered here, namely the sphere.

The biasing potential $\mathcal{B}_k(\mathbf{Q})$ corresponding to a sphere \mathcal{S}_k associated with a state k is defined by the following parameters: a sphere center \mathbf{Q}_k , a radius R_k , a restraining force constant c_k , a number of radial grid points $\Gamma_k + 1$, and a memory force-constant vector $\mathbf{M}_k = \{M_{k,i}, i = 0, \dots, \Gamma_k\}$. The corresponding expression is (for $k = 1, \dots, K$)

$$\mathcal{B}_k(\mathbf{Q}) = \begin{cases} M_{k,\Gamma_k} + \frac{1}{2}c_k(r_k - R_k)^2 & \text{if } r_k \geq R_k \\ \sum_{i=0}^{\Gamma_k} M_{k,i}\gamma(d_{k,i}) & \text{if } r_k < R_k \end{cases} \quad (1)$$

where the function γ is defined as³⁷

$$\gamma(x) = H(1 - |x|)(1 - 3x^2 + 2|x^3|) \quad (2)$$

H being the Heaviside step function, and the quantities r_k and $d_{k,i}$ depend on \mathbf{Q} as

$$r_k = \|\mathbf{Q} - \mathbf{Q}_k\| \quad (3)$$

and

$$d_{k,i} = \Gamma_k R_k^{-1} r_k - i \quad (4)$$

Within the sphere ($r_k < R_k$), the memory vector \mathbf{M}_k permits to enforce a radially dependent potential of arbitrary form

(with an approximate resolution $\Gamma_k^{-1}R_k$), expressed as a weighted sum of $\Gamma_k + 1$ repulsive local functions γ . Outside the sphere ($r_k \geq R_k$), the potential is changed to an attractive half-harmonic restraint. It is easily verified that the biasing potential \mathcal{B}_k defined by eq 1 is continuous and differentiable.³⁷

The extension to nonspherical centered volumes (e.g., ellipsoids or polyhedra) is in principle straightforward and requires the generalization of the radius R_k to a function $R_k(\mathbf{Q})$ accounting for the center–surface distance in the different directions. The “radial” grid points will then map to scaled versions of this surface rather than to spheres. In any case, the internal dimensionality of the biasing potential is one, which implies a limited memory cost. Extending this dimensionality so as to include some memory-based directional dependence (still of limited dimensionality) into the biasing potential is certainly feasible, but should not be required if the reduced conformational subspace has been chosen appropriately, that is, if it encompasses enough variables to define all states as distinct regions. A carefully adjusted one-dimensional memory may be used to obtain a biasing potential enforcing a homogeneous radial sampling of the centered volume, but this potential will generally not lead to a homogeneous sampling of the multidimensional volume itself, the directional (nonradial) dimensions remaining unbiased. Note also that such a potential will guarantee that the center of the volume is sampled.

2.3. Definition of a Set of Conformational Paths. The third step (C) in the B&S-LEUS approach relies on the definition of a set of L conformational paths connecting the centers of the volumes representing the K states. Only the simplest possible type of path will be considered here, namely the line (or, more precisely, the line segment). The numbering of the corresponding biasing potentials will start at $K + 1$.

The biasing potential $\mathcal{B}_l(\mathbf{Q})$ corresponding to a line \mathcal{L}_l is defined by the following parameters: a starting point \mathbf{Q}_l , an ending point \mathbf{Q}'_l , a (double) width W_l , a restraining force constant c_l , a number of longitudinal grid points $\Gamma_l + 1$, and a memory force-constant vector $\mathbf{M}_l = \{M_{l,i}, i = 0, \dots, \Gamma_l\}$. The corresponding expression is (for $l = K + 1, \dots, K + L$)

$$\mathcal{B}_l(\mathbf{Q}) = \begin{cases} M_{l,0} + \frac{1}{2}c_l H(r_l - W_l)(r_l - W_l)^2 & \text{if } u_l \leq 0 \\ M_{l,\Gamma_l} + \frac{1}{2}c_l H(r'_l - W_l)(r'_l - W_l)^2 & \text{if } u_l \geq U_l \\ \sum_{i=0}^{\Gamma_l} [M_{l,i} + \frac{1}{2}c_l H(p_l - W_l)(p_l - W_l)^2] \gamma(d_{l,i}) & \text{if } 0 < u_l < U_l \end{cases} \quad (5)$$

where the function γ is defined by eq 2, U_l is the line length

$$U_l = \|\mathbf{Q}'_l - \mathbf{Q}_l\| \quad (6)$$

and the quantities u_l , p_l , r_l , r'_l , and $d_{l,i}$ depend on \mathbf{Q} as

$$u_l = U_l^{-1}(\mathbf{Q}'_l - \mathbf{Q}_l)^T(\mathbf{Q} - \mathbf{Q}_l) \quad (7)$$

$$p_l = \|(\mathbf{Q} - \mathbf{Q}_l) - U_l^{-1}u_l(\mathbf{Q}'_l - \mathbf{Q}_l)\| \quad (8)$$

$$r_l = \|\mathbf{Q} - \mathbf{Q}_l\| \quad (9)$$

$$r'_l = \|\mathbf{Q} - \mathbf{Q}'_l\| \quad (10)$$

and

$$d_{l,i} = \Gamma_l U_l^{-1}u_l - i \quad (11)$$

v^T indicating the transpose of a vector v . Note that due to the identity³⁷

$$\sum_{i=0}^{\Gamma_l} \gamma(d_{l,i}) = 1 \quad \text{if } 0 < u_l < U_l \quad (12)$$

the third conditional statement in eq 5 could in principle be simplified to

$$\mathcal{B}_l(\mathbf{Q}) = \frac{1}{2}c_l H(p_l - W_l)(p_l - W_l)^2 + \sum_{i=0}^{\Gamma_l} M_{l,i} \gamma(d_{l,i}) \quad \text{if } 0 < u_l < U_l \quad (13)$$

This simplification was not undertaken, so as to allow for a possible generalization to displaced lines and lines with longitudinally dependent widths or force constants (see below). The quantity u_l represents the longitudinal distance between the starting point of the line and the current point \mathbf{Q} , while the parameter p_l represents the corresponding transverse (perpendicular) distance. Within the line ($0 < u_l < U_l$), the memory vector \mathbf{M}_l permits to enforce a longitudinally dependent potential of arbitrary form (with an approximate resolution $\Gamma_l^{-1}U_l$), expressed as a weighted sum of $\Gamma_l + 1$ repulsive local functions γ , and applied together with a transverse attractive flat-bottom (width W_l) half-harmonic restraining potential. Outside the line, that is, when going past its two terminal points in terms of longitudinal distance ($u_l \leq 0$ or $u_l \geq U_l$), the potential is changed to an attractive flat-bottom (width W_l) half-harmonic restraint depending on the distance to the corresponding end point. It is easily verified that the biasing potential \mathcal{B}_l defined by eq 5 is continuous and differentiable.

The extension to nonlinear paths (e.g., arbitrary curves) is in principle possible and requires the generalization of the end points \mathbf{Q}_l and \mathbf{Q}'_l to a parametric path $\mathbf{Q}_l(u)$, where $\mathbf{Q}_l(0)$ is the starting point and $\mathbf{Q}_l(U_l)$ is the ending point, U_l being the path length. A simpler variant of this approach, the displaced line, involves modifying a normal line by allocating to all nonterminal grid points an offset coordinate $\Delta\mathbf{Q}_{l,i}$ perpendicular to the line (i.e., with $\Delta\mathbf{Q}_{l,i}^T(\mathbf{Q}'_l - \mathbf{Q}_l) = 0$ and $\Delta\mathbf{Q}_{l,0} = \Delta\mathbf{Q}_{l,\Gamma_l} = 0$), and replacing p_l in eq 5 by

$$p_{l,i} = \|(\mathbf{Q} - \mathbf{Q}_l) - U_l^{-1}u_l(\mathbf{Q}'_l - \mathbf{Q}_l) - \Delta\mathbf{Q}_{l,i}\| \quad (14)$$

Another possible variant involves the use of longitudinally dependent line widths or/and restraining force constants, that is, the replacement of c_l and W_l in eq 5 by corresponding grid-point dependent quantities $c_{l,i}$ and $W_{l,i}$. In any case, the internal dimensionality of the biasing potential is one, which implies a limited memory cost. Extending this dimensionality

so as to include some memory-based transverse dependence (still of limited dimensionality) into the biasing potential is certainly feasible, but not required in practice, because paths are only meant to promote transitions between states and typically located in irrelevant regions of the reduced conformational subspace. A carefully adjusted one-dimensional memory may be used to obtain a biasing potential enforcing a homogeneous longitudinal sampling of the path, but this potential will generally not lead to a homogeneous transverse sampling. Note also that, in contrast to the centered volume case, such a potential will not automatically guarantee that the two end points of the path are sampled, although this will generally be the case in practice if the line width is sufficiently small (or can be enforced otherwise by decreasing the line width close to the end points).

In principle, the L paths will be chosen to connect pairs among the K centered volumes defining the states. This must be done in such a way that all states are connected to each other via at least one path or succession thereof. The minimum number of paths is thus $L = K - 1$ (maximum-spanning tree), but it may be advantageous in terms of convergence properties to include additional (redundant) paths. Note that the end points of the paths must be identical to the centers of the states (e.g., they should not connect to the periphery of the centered volumes). This is essential because independent biasing potentials leading to a homogeneous radial sampling of the centered volumes and longitudinal sampling of the paths can only guarantee that these specific points are sampled (for lines, assuming sufficiently small line widths, at least at the end points).

2.4. Unification of the Biasing Potentials. The fourth step (D) in the B&S-LEUS approach relies on the unification of the biasing potentials associated with the $M = K + L$ centered volumes and paths into a single biasing potential according to the EDS procedure.⁷⁸ For the ease of notation, these objects have been given the generic notation B_m , where the index m ranges from 1 to $M = K + L$ (1, ..., K for the centered volumes, $K + 1$, ..., M for the paths). The various LEUS potentials are combined following the EDS principle as

$$\mathcal{U}_{\text{bias}}(\mathbf{r}; \mathbf{M}) = -\frac{1}{\beta s} \ln \left(\sum_{m=1}^M \exp[-\beta s B_m(\mathbf{Q}(\mathbf{r}))] \right) \quad (15)$$

where $\beta = (k_B T)^{-1}$, k_B being Boltzmann's constant and T the absolute temperature, \mathbf{r} represents the system configuration (Cartesian coordinates of all particles), $\mathbf{Q}(\mathbf{r})$ the corresponding representative point in the reduced subspace, s a (positive) smoothing parameter, and \mathbf{M} the joint memories of the M objects, that is, a vector containing $N_M = \sum_{m=1}^M (\Gamma_m + 1)$ elements.

Qualitatively speaking, the exponential weighting in eq 15 ensures that the combined biasing potential $\mathcal{U}_{\text{bias}}$ is low in the regions of the conformational subspace where any of the B_m is low, and high in the regions where all of the B_m are high. For the ease of reference, the subvolume of the reduced conformational subspace where any of the B_m is low, that is, the union of all centered volumes and paths, will be referred to as the active subspace. In the absence of

memory ($\mathbf{M} = \mathbf{0}$) and assuming a hypothetical (physical-system) dynamics leading to a homogeneous sampling in terms of \mathbf{Q} , the introduction of $\mathcal{U}_{\text{bias}}$ will effectively restrict the sampling to the active subspace by action of the individual restraining potentials within the M objects, with a slight bias toward regions where multiple objects overlap. In the presence of a memory ($\mathbf{M} \neq \mathbf{0}$) and considering a real molecular system (physical potential energy function affecting the sampling in terms of \mathbf{r}), the regions sampled will be determined by an interplay between five factors: (i) a strong bias toward the active subspace induced by the restraints; (ii) a slight bias toward the regions where multiple objects overlap; (iii) a physical bias in terms of \mathbf{r} ; (iv) a bias related to the Jacobian of the $\mathbf{Q}(\mathbf{r})$ transformation; and (v) a memory-based bias within the objects. The goal of the LE build-up phase will be to adjust the memory so that the four latter types of biases cancel out, leading, in conjunction with the first bias, to a nearly homogeneous (radial within the centered volumes, longitudinal within the paths) sampling, restricted to the active subspace. Note that the requirements of homogeneous sampling within all single objects may in some cases be conflicting at the level of the unified biasing potential, in regions where multiple objects overlap. However, these homogeneity violations are expected to be marginal, especially in high-dimensional cases.

The forces derived from U_{bias} in eq 15 are given by

$$\mathbf{F}_{\text{bias}}(\mathbf{r}) = -\frac{\partial \mathcal{U}_{\text{bias}}(\mathbf{r}; \mathbf{M})}{\partial \mathbf{r}} = -\sum_{m=1}^M w_m(\mathbf{Q}) \frac{d\mathcal{B}_m(\mathbf{Q})}{d\mathbf{Q}} \frac{\partial \mathbf{Q}}{\partial \mathbf{r}} \quad (16)$$

where

$$w_m(\mathbf{Q}) = \frac{\exp[-\beta s \mathcal{B}_m(\mathbf{Q})]}{\sum_{m=1}^M \exp[-\beta s \mathcal{B}_m(\mathbf{Q})]} \quad (17)$$

can be interpreted as measuring the relative influence (weight) of a single-object biasing potential m on the dynamics of the system in a conformation \mathbf{Q} . Note that the forces defined by eq 16 are nonsingular, because $\mathcal{B}_m(\mathbf{Q})$ and $\mathbf{Q}(\mathbf{r})$ are both differentiable functions of their arguments. The parameter s in eqs 15 and 17 will affect the extent to which differences between the M single-object biasing potentials are “smoothed out” in $\mathcal{U}_{\text{bias}}$. Because the paths connect to the states at their centers, this parameter will have little influence on the biasing potential within the active subspace. On the other hand, it will affect the “sharpness” with which the restraining potentials define this subspace, a high value indicating a “sharper” combination (with the risk of occurrence of high restraining forces) and a low value a more “fuzzy” combination (with the risk of unnecessarily increasing the irrelevant volume). Because this parameter is nevertheless expected to have a minor overall influence on the B&S-LEUS scheme when selected within reasonable bounds, it was simply set to $s = 1$ in the present study.

2.5. LE Build-Up Phase. The fifth step (E) of the B&S-LEUS procedure (LE build-up phase, duration t_{LE}) relies on the optimization of the memory, leading to a biasing potential

enabling nearly uniform sampling (radially within the centered volumes, longitudinally within the paths) of the active subspace. In this phase, the (time-dependent) potential-energy function used for (thermostatted) MD searching is written

$$\mathcal{U}(\mathbf{r}; \mathbf{M}(t)) = \mathcal{U}_{\text{phys}}(\mathbf{r}) + \mathcal{U}_{\text{bias}}(\mathbf{r}; \mathbf{M}(t)) \quad (18)$$

where $\mathcal{U}_{\text{phys}}$ is the physical potential-energy function (force field) and $\mathcal{U}_{\text{bias}}$ is given by eq 15. The updating scheme for the memory $\mathbf{M}(t)$ relies on the equation

$$\mathbf{M}_{m,i}(t + \Delta t) = \mathbf{M}_{m,i}(t) + k_{\text{LE}} f_{\text{LE}}^{I_{\mathcal{R}}(\gamma_{\text{LE}}, n_{\text{LE}})} j_{m,i}(\mathbf{Q}) h_{m,i}(\mathbf{Q}) w_m(\mathbf{Q}) \quad (19)$$

where $\mathbf{M}_{m,i}$ is the memory associated with grid point i of object m , $\mathbf{Q} = \mathbf{Q}(\mathbf{r}(t))$, Δt is the simulation time step, k_{LE} is the basis force-constant increment, f_{LE} is a force-constant reduction factor, $I_{\mathcal{R}}$ is a force-constant reduction counter, associated with a defined conformational region \mathcal{R} , γ_{LE} is a local visiting cutoff (real), n_{LE} is a global visiting cutoff (integer), $j_{m,i}$ is a distribution-alteration function, $h_{m,i}$ is a grid-assignment function, and w_m is the weight defined by eq 17. In the absence of prior knowledge concerning the form on the free-energy hypersurface, the memory will typically be initiated to $\mathbf{M}(0) = \mathbf{0}$. The different factors involved in eq 19 are explained below.

The grid-assignment function evaluates to one for a single grid point in each of the single-object biasing potentials (and to zero for all other grid points), namely the grid point i in object m that is (radially for centered volumes, longitudinally for paths) closest to \mathbf{Q} . For a sphere k , one has ($k = 1 \dots K$)

$$h_{k,i}(\mathbf{Q}) = \begin{cases} \delta_{i,\Gamma_k} & \text{if } r_k \geq R_k \\ \delta_{i,\text{NINT}(\Gamma_k R_k^{-1} r_k)} & \text{if } r_k < R_k \end{cases} \quad (20)$$

where δ is the Kronecker symbol and the function NINT returns the nearest integer to a real number. For a line l , one has ($l = K + 1, \dots, M$)

$$h_{k,i}(\mathbf{Q}) = \begin{cases} \delta_{i,0} & \text{if } u_l \leq 0 \\ \delta_{i,\Gamma_l} & \text{if } u_l \geq U_l \\ \delta_{i,\text{NINT}(\Gamma_l U_l^{-1} u_l)} & \text{if } 0 < u_l < U_l \end{cases} \quad (21)$$

As a result, the build-up always affects one and only one grid point in each of the M single-object memories. However, the presence of the weight factor w_m in eq 19 ensures that the build-up is only significant within the objects encompassing or closest to point \mathbf{Q} (note that the sum of w_m over all objects is one).

The distribution-alteration function is generally set to

$$j_{m,i}(\mathbf{Q}) = 1 \quad (22)$$

leading to a nearly homogeneous sampling (radially within the centered volumes, longitudinally within the paths) of the active subspace. However, this function may be used to enforce deviations from this homogeneous sampling. As a simple example, one may observe that the volume of relevant conformational subspace accounted for by a radial grid point i within a sphere k (distance $\Gamma_k^{-1} i R_k$ from the center) increases

with $(i_k + 1/2)^{N-1}$ (Jacobian factor), where N is the subspace dimensionality. One may then decide to bias the sampling of the sphere toward its periphery, which can be achieved by setting for all spheres k ($k = 1, \dots, M$)

$$j_{k,i}(\mathcal{Q}) = \frac{\left(i + \frac{1}{2}\right)^{1-N}}{\sum_{j=0}^{\Gamma_k} \left(j + \frac{1}{2}\right)^{1-N}} \quad (23)$$

Unless otherwise specified, eq 22 (rather than eq 23) was employed in the present study.

The force-constant reduction factor can be used in the context of an iterative procedure to progressively decrease the build-up rate during the searching phase. As noted previously by other authors,^{86,106} a high build-up rate is desired in the early stage of the searching, where the deep free-energy basins have to be “filled up” coarsely (i.e., without wasting computer time), while a low build-up rate (near-equilibrium situation) is preferable in the later stage, where the remaining shallower free-energy wiggles have to be “leveled off” (so as to produce a close-to-optimal biasing potential). This can be achieved by a progressive reduction of the build-up rate, enforced in eq 19 by using $f_{LE} < 1$ along with a force-constant reduction counter $I_{\mathcal{R}}$ progressively increasing with time (the choice $f_{LE} = 1$ switches off the force-reduction procedure). In the B&S-LEUS algorithm, the force-reduction procedure is associated with a region \mathcal{R} within the active subspace, defined by a specific collection of grid points. The reduction counter $I_{\mathcal{R}}$ is propagated in time according to the following procedure. $I_{\mathcal{R}}(t)$ as well as an auxiliary counter $N_c(t)$ are set to 0 at $t = 0$. An auxiliary memory $A(t)$ is also set to $\mathbf{0}$ at $t = 0$ and propagated in time according to the equation

$$A_{m,i}(t + \Delta t) = A_{m,i}(t) + w_m(\mathcal{Q})h_{m,i}(\mathcal{Q}) \quad (24)$$

When $A_{m,i}(t)$ exceeds a specified local visiting cutoff γ_{LE} for all grid points $(m,i) \in \mathcal{R}$, the auxiliary counter is increased by one and the auxiliary memory reset to zero. When the auxiliary counter exceeds a global visiting cutoff n_{LE} , $I_{\mathcal{R}}$ is increased by one and the auxiliary counter reset to zero. Two possible (reasonable) choices for \mathcal{R} are: (i) the $i = 0$ (central) grid points of all centered volumes k ($k = 1, \dots, K$), a choice that will be noted $\mathcal{R} = \mathcal{C}$; and (ii) all grid points i of all objects m ($m = 1, \dots, M$), a choice that will be referred to as $\mathcal{R} = \mathcal{A}$. Possible (reasonable) choices for the parameters γ_{LE} and n_{LE} are 1.0 and 2, respectively. The reasoning behind the present force-constant reduction scheme (assuming $\gamma_{LE} = 1.0$ and $n_{LE} = 2$) is that when all grid points of \mathcal{R} have undergone an “effective” number of visits (auxiliary memory, i.e., based on the w_m weights) of one, it is still possible that the “flattened” free-energy hypersurface retains an overall “slope”. However, when all of these points have undergone an “effective” number of visits of one for the second time, even the points that were “uphill” have been revisited. When this condition is met, it becomes advantageous to reduce the build-up rate by incrementing $I_{\mathcal{R}}$, which in effect scales this rate by a factor f_{LE} .

Finally, the constant k_{LE} in eq 19 represents the basic force-constant increment (units of energy) and determines the initial rate of the build-up. Note that the above force-reduction procedure also presents the advantage of permitting a convergence assessment of the build-up phase, by monitoring the time evolution of $I_{\mathcal{R}}$. The build-up phase can, for example, be terminated whenever $I_{\mathcal{R}}$ reaches a threshold value $I_{\mathcal{R}}^{\max}$. In this case, the procedure guarantees that all grid points of \mathcal{R} have undergone an “effective” number of visits of at least $n_{LE}\gamma_{LE}I_{\mathcal{R}}^{\max}$, while the energetic resolution of the biasing potential is of the order of $f_{LE}^{\max}k_{LE}$. Alternatively, the termination may be based on the time interval separating successive incrementations of $I_{\mathcal{R}}$. In the initial stage of the build-up, the diffusion of the system within the active subspace will be accelerated (hill surfing). However, as the free-energy hypersurface becomes increasingly “flat” and the build-up rate is decreased, this diffusion will progressively slow down toward a “natural” regime (as determined by the physical system after removal of the free-energy bias). Thus, the force-reduction procedure could also be terminated when the interval separating successive increments of $I_{\mathcal{R}}$ has increased and leveled off to an approximately constant time.

Two specific features of the B&S-LEUS scheme make it particularly robust with respect to the details of the build-up protocol and well-suited for an automatic force-reduction procedure, as compared to the standard LEUS scheme³⁷ and related approaches.^{52,54} First, in the B&S-LEUS scheme, the grid points are defined a priori, and it is known that all of the corresponding grid cells must have been sufficiently sampled during a successful build-up phase. In contrast, in the standard LEUS scheme, new grid points are steadily added, and because the total number of possible grid points within the reduced subspace is typically extremely large, the build-up is stopped long before all of them have been visited. As a result, it can never be guaranteed that all of the relevant grid points have actually been visited, and an inappropriate build-up procedure (e.g., too high build-up rate or too fast force-reduction procedure) may lead to the omission of important conformational regions. Second, in the B&S-LEUS scheme, the effective volume of the active subspace is unaffected by the build-up. The reason is that the confinement potentials are “attached” to the surface of the different objects (centered volumes or paths, see eqs 1 and 5). As a result, they “rise” simultaneously with the memory-based component of the biasing potential. This can equivalently be seen by observing that the addition of an arbitrary vector with identical components to the memory \mathbf{M} in eq 15 only changes $\mathcal{U}_{\text{bias}}$ by a constant, that is, does not affect the dynamics. In contrast, in the standard LEUS scheme, the “rise” of the memory-based potential slowly expands the accessed conformational volume, predominantly including irrelevant (high free energy) regions.

Similarly to the original LEUS method,¹⁵ a successful LE build-up phase will generate a biasing potential that is (approximately) equal to the negative of the free-energy function $G(\mathcal{Q})$ in the active subspace, that is

$$\mathcal{U}_{\text{bias}}(\mathbf{r}; \mathbf{M}(t_{LE})) \approx -G(\mathcal{Q}) \quad (25)$$

There is a slight difference in the interpretation of this approximate equality within the plain LEUS and B&S-LEUS approaches. In plain LEUS, this result will only be valid up to a certain free-energy threshold above the lowest free-energy point encountered, the latter threshold depending on the duration and rate of the build-up. However, the presence of regions with a free energy lower than this threshold, but separated by high barriers from the free-energy basin that has been searched, cannot be ruled out. In the B&S-LEUS scheme, the states to be sampled are known in advance, and it can be verified explicitly that all of them have been visited (e.g., definition of \mathcal{R} in the force-reduction procedure). However, it still cannot be guaranteed that eq 25 holds everywhere within the active subspace, because the sampling within the states is only enhanced along one (radial) degree of freedom, leaving room for barriers along the $N - 1$ remaining ones. In other words, $\mathcal{U}_{\text{bias}}$ may be characteristic of a limited subregion of each state, while G characterizes the entire state.

As discussed elsewhere,¹⁵ the approximate equality in eq 25 is in general not sufficient for an accurate evaluation of the relative free energies of the conformational states, which requires a subsequent sampling phase.

2.6. US Sampling Phase. The sixth step (F) of the B&S-LEUS procedure (US sampling phase, duration t_{US}) relies on the generation of a biased ensemble of configurations, using the biasing potential preoptimized during the LE build-up phase. In this phase, the (time-independent) potential-energy function used for (thermostatted) MD sampling is written

$$\mathcal{U}(\mathbf{r};\mathbf{M}) = \mathcal{U}_{\text{phys}}(\mathbf{r}) + \mathcal{U}_{\text{bias}}(\mathbf{r};\mathbf{M}) \quad (26)$$

where \mathbf{M} is from here on a short notation for $\mathbf{M}(t_{\text{LE}})$. Due to eqs 15 and 25, the biased sampling during this phase should be approximately homogeneous (radially within the centered volumes, longitudinally along the paths) within the active subspace. In particular, all states should be sampled with equal populations (assuming identical radii and numbers of grid points for all spheres). In addition, due to the presence of the connecting paths, diffusive interconversion transitions should occur frequently between these states. As a result, an accurate evaluation of their relative free energies becomes possible.

2.7. Reweighting and State Assignment. The seventh step (G) of the B&S-LEUS procedure (reweighting and state assignment) involves the postprocessing of the data accumulated during the sampling phase, so as to calculate the relative free energies of the states in the physical ensemble. For each state k , the free energy can be written (for $k = 1, \dots, K$):

$$G_k = -\beta^{-1} \ln \langle \exp[\beta \mathcal{U}_{\text{bias}}(\mathbf{r};\mathbf{M})] \rangle_{\mathcal{Q}(\mathbf{r}) \in \mathcal{S}_k'} + C_G \quad (27)$$

where C_G is an arbitrary offset constant (typically chosen so that $G_k = 0$ for the lowest free-energy state) and $\langle \dots \rangle_{\mathcal{Q}(\mathbf{r}) \in \mathcal{S}_k'}$ denotes ensemble (trajectory) averaging over the biased ensemble (sampling phase), restricted to conformations belonging to state k . The symbol \mathcal{S}_k' has been used rather than \mathcal{S}_k to underline the fact that the regions used to assign

the states need not necessarily be exactly identical to the centered volumes involved in the construction of the biasing potential. Although the calculation of relative free energies is the main concern of the present work, it should be stressed that reweighting formulas analogous to eq 24 can be formulated for any other type of thermodynamic (e.g., enthalpy, entropy, heat capacity, or volume) or structural (ensemble average of a given instantaneous observable) quantity. Note that the reweighting requires the quantity $\mathcal{U}_{\text{bias}}$ to be stored along with each successive frame along the trajectory.

2.8. Additional Remarks. The B&S-LEUS approach involves “physical” parameters, namely the definition of the reduced conformational subspace and the choice of the centered volumes (central conformation and radius) associated with the states. In favorable situations, the free-energy hypersurface will involve well-defined basins around the central conformations selected to represent the different states. In this case, the results should be relatively insensitive to the selected radii, provided that they are chosen large enough to encompass the low free-energy regions of the different states. In practice, however, basins may not be centered at the selected reference conformations, for example, due to an inaccurate choice of these conformations or to the approximate force-field representation of the system. Consequently, the sensitivity of the calculated relative free energies to these parameters should be investigated. On the long run, a method where the centers and radii (and possibly shapes) of the centered volumes defining the states are refined adaptively could be envisioned.

The B&S-LEUS approach also involves “numerical” parameters, namely the choice of a specific set of paths along with the corresponding path widths, the restraining force constants, the numbers of grid points used for the centered volumes and paths, the EDS smoothing parameter, the various parameters of the build-up procedure, the build-up time t_{LE} , and the sampling time t_{US} . The sensitivity of the calculated relative free energies to these parameters should also be investigated.

Although the B&S-LEUS approach has been formulated here in terms of a common reduced space \mathcal{Q} for all objects (e.g., set of conformationally relevant dihedral angles), this requirement is actually not critical to the method. Of particular interest would be the definition on centered volumes based on the root-mean-square atomic positional deviation from corresponding reference structures. The paths would then represent arbitrary configurational pathways interconverting the reference structures of the connected states (these could be constructed, e.g., via reaction-path approaches^{107–110} or targeted molecular dynamics^{28,111,112}). The latter variant would be practically extremely relevant, providing a direct connection (via simulation and statistical mechanics) between three-dimensional structures (as available experimentally from X-ray or NMR determinations) and free-energy differences (between conformational states defined as their configurational neighborhoods).

As a final remark, it should be kept in mind that the B&S-LEUS approach will only be able to handle conformational states of “limited” extents (i.e., amenable to sufficient

Table 1. Characteristics of the Conformational Objects (Spheres or/and Lines) Defining the B&S-LEUS Biasing Potentials in the Different Simulations^a

system	M	object	Q_m	Q'_m	R_m	W_m	c_m [kJ mol ⁻¹]	$\Gamma_m + 1$
A ₁	0							
A ₂	1	\mathcal{S}_1	(180,180)		45		0.5	10
A ₃	1	\mathcal{S}_1	(180,180)		45		0.5	10
A ₄	1	\mathcal{L}_1	(60,60)	(300,300)		10	0.1	20
A ₅	1	\mathcal{L}_1	(60,60)	(300,300)		b	0.1	20
A ₆	1	\mathcal{L}_1^c	(60,60)	(300,300)		10	0.1	20
A ₇	2	\mathcal{L}_1	(60,60)	(300,300)		10	0.5	20
		\mathcal{L}_2	(300,60)	(60,300)		10	0.5	20
A ₈	6	\mathcal{L}_1	(60,60)	(60,300)		10	0.5	20
		\mathcal{L}_2	(60,60)	(300,60)		10	0.5	20
		\mathcal{L}_3	(60,60)	(300,300)		10	0.5	20
		\mathcal{L}_4	(60,300)	(300,300)		10	0.5	20
		\mathcal{L}_5	(300,60)	(60,300)		10	0.5	20
		\mathcal{L}_6	(300,60)	(300,300)		10	0.5	20
A ₉	8	\mathcal{L}_1	(50,60)	(70,300)		10	0.5	20
		\mathcal{L}_2	(160,60)	(180,300)		10	0.5	20
		\mathcal{L}_3	(230,60)	(250,300)		10	0.5	20
		\mathcal{L}_4	(290,60)	(310,300)		10	0.5	20
		\mathcal{L}_5	(50,60)	(90,60)		10	0.5	20
		\mathcal{L}_6	(60,180)	(100,180)		10	0.5	20
		\mathcal{L}_7	(70,300)	(250,300)		10	0.5	20
		\mathcal{L}_8	(240,180)	(300,180)		10	0.5	20
A ₁₀	6	\mathcal{S}_1	(60,60)		45		0.5	10
		\mathcal{S}_2	(60,300)		45		0.5	10
		\mathcal{S}_3	(300,60)		45		0.5	10
		\mathcal{S}_4	(300,300)		45		0.5	10
		\mathcal{L}_1	(60,60)	(300,300)		10	0.5	20
		\mathcal{L}_2	(300,60)	(60,300)		10	0.5	20
P ₁	4	\mathcal{S}_1	(-130) ₉		115.5		0.02	15
		\mathcal{S}_2	(-105) ₉		115.5		0.02	15
		\mathcal{S}_3	(-75) ₉		115.5		0.02	15
		\mathcal{L}_1	(-130) ₉	(-75) ₉		8.25	0.02	21
P ₂	1	\mathcal{L}_1	(-150) ₉	(-50) ₉		100	0.02	21
H ₁	63	\mathcal{S}_{1-32}	d		50.0		1.0	6
		\mathcal{L}_{1-31}	e	e		10.0	1.0	14
H ₂	112	\mathcal{S}_{1-32}	d		50.0		1.0	6
		\mathcal{L}_{1-80}	f	f		10.0	1.0	14

^a The blocked alanine mono-peptide (A₁–A₁₀), unblocked polyalanine decapeptide (P₁ and P₂), and artificial hexopyranose (H₁ and H₂) systems, as well as the corresponding relevant conformational subspaces (two, nine, and seven dimensions, respectively) are described in sections 3.1–3.3. For each object m ($m = 1, \dots, M$, with $M = K + L$, where K is the number of spheres and L the number of lines), the following quantities are reported: center (sphere) or starting point (line) Q_m , ending point (line only) Q'_m , radius (sphere only) R_m , width (line only) W_m , restraining force constant c_m , and number of radial (sphere) or longitudinal (line) grid-points $\Gamma_m + 1$. For the decapeptide system, Q_m and Q'_m are nine-dimensional vectors with identical components, indicated as $(Q)_9$, where Q is the corresponding common value. The corresponding B&S-LEUS protocol parameters are listed in Table 2. ^b Longitudinally dependent width, as described in section 2.3, using $W_1 = 10$ (6, 5, 5, 4, 4, 3, 3, 2, 2, 1, 1, 2, 2, 3, 3, 4, 4, 5, 5, 6). ^c Displaced line, as described in section 2.3, using $\Delta Q_1 = 5(2^{1/2})(-1, 1)d$, where $d = (0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 9, 8, 7, 6, 5, 4, 3, 2, 1, 0)$. ^d Centered at ideal values appropriate for the 32 isomers, as described in section 3.3. ^e Maximum-spanning tree as described in section 3.3 and illustrated in Figure 3b. ^f Redundant spanning tree (Manhattan metric) as described in section 3.3.

sampling on the time scale accessible to the simulation). For example, an extension of the scheme to the determination of the folding free energy of peptides (or proteins) appears difficult, because a proper sampling of the unfolded state would require longer time scales than currently feasible, except for the smallest systems.

3. Computational Details

Three test systems were considered to assess the performance of the B&S-LEUS scheme. This section provides the computational details concerning these systems and the corresponding simulations. In the three cases, the reduced conformational subspace is defined in terms of a set of angular coordinates. The corresponding reference values σ_n (section 2.1) are systematically set to $\sigma = 1$ degree, leading to unitless variables. This convention is applied throughout,

but unitless quantities X (e.g., coordinates Q_n , sphere radii R_k , and line widths W_k) are sometimes reported as σX in units of degrees for the clarity of the discussion. Note that, although the variables Q_n are not “refolded” to a reference period during the calculations, this “refolding” is actually performed in the displayed figures.

3.1. Blocked Alanine Mono-peptide. A solvated (blocked) alanine mono-peptide was used as a toy system to illustrate the versatility of the B&S-LEUS method in promoting the sampling of arbitrary regions of the Ramachandran map. This mono-peptide consists of an N-acetylated (ac) and C-methylamidated (am) alanine residue (see Figure 1 in ref 37) and was simulated in the presence of $N_{H_2O} = 1300$ water molecules (see section 3.4 for details). The two-dimensional reduced conformational subspace was defined by the variables $Q_1 = \sigma^{-1}\phi$ and $Q_2 = \sigma^{-1}\psi$, where ϕ and ψ are the

Table 2. Parameters of the B&S-LEUS Protocol Used in the Different Simulations^a

system	<i>s</i>	k_{LE} [kJ mol ⁻¹]	\mathcal{R}	f_{LE}	γ_{LE}	n_{LE}	t_{LE} [ns]	I_{R}^{max}	t_{US} [ns]
A ₁ ^b		2×10^{-3}		1.0			15		50
A ₂		5×10^{-5}		1.0			5		1
A ₃ ^c		5×10^{-5}		1.0			5		1
A ₄		1×10^{-4}		1.0			5		1
A ₅		1×10^{-4}		1.0			5		1
A ₆		1×10^{-4}		1.0			5		1
A ₇	1.0	2×10^{-4}		1.0			10		1
A ₈	1.0	5×10^{-4}		1.0			20		5
A ₉	1.0	5×10^{-4}		1.0			20		5
A ₁₀	1.0	2×10^{-4}		1.0			20		5
P ₁	1.0	1×10^{-3}	\mathcal{B}	0.5	1	2	50	6	2×50^d
P ₂	1.0	5×10^{-3}	\mathcal{A}	0.5	1	2	20	7	2×50^d
H ₁	1.0	5×10^{-2}	\mathcal{B}	0.5	1	2	100	6	100
H ₂	1.0	5×10^{-2}	\mathcal{A}	0.5	1	2	100	8	100

^a The blocked alanine mono-peptide (A₁–A₁₀), unblocked polyanaline decapeptide (P₁ and P₂), and artificial hexopyranose (H₁ and H₂) systems, as well as the corresponding relevant conformational subspaces (two, nine, and seven dimensions, respectively) are described in sections 3.1–3.3. The different parameters are defined in sections 2.4–2.6. Unless noted otherwise, the distribution-alteration function was set to one (eq 22). The indicated parameters are the basis force constant k_{LE} , the EDS smoothing parameter s , the conformational region \mathcal{R} involved in the force-constant-reduction procedure (\mathcal{B} : union of all central grid points of the conformational spheres, \mathcal{A} : union of all grid points of all of the conformational lines and spheres), the force-constant reduction factor f_{LE} , the local visiting cutoff γ_{LE} , the global visiting cutoff n_{LE} , the duration of the LE-build-up phase t_{LE} , the value of the force-reduction counter at the end of the build-up phase I_{R}^{max} , and the duration of the US sampling phase t_{US} . The characteristics of the associated conformational objects (spheres or/and lines) are listed in Table 1. ^b Plain LEUS simulation from ref 37. ^c Using the distribution-alteration function of eq 23. ^d Two sampling trajectories of 50 ns each were concatenated.

dihedral angles corresponding to the atom sequences C_{ac}–N–C_α–C_{CO} and N–C_α–C_{CO}–N_{am}, respectively. Ten different forms of biasing potential were considered, labeled A₁–A₁₀, the corresponding parameters being reported in Tables 1 and 2.

3.2. Polyanaline Decapeptide. A solvated (unblocked) polyanaline decapeptide was used as an application of the B&S-LEUS scheme to the determination of the relative free energies of different types of helices (π , α , and 3_{10}), based on a single simulation. This oligopeptide consists of a sequence of 10 alanine residues with free termini (unprotonated amine and protonated carboxylic acid) and was simulated in the presence of $N_{H_2O} = 3300$ water molecules (see section 3.4 for the simulation details). The nine-dimensional reduced conformational subspace was defined by the variables $Q_n = \sigma^{-1}\chi_n$ ($n = 1, \dots, 9$), where χ_n is the sum of the two dihedral angles ϕ_{n+1} and ψ_n encompassing the successive peptide bonds¹¹³ (see Figure 1 in ref 37). Note that these angles do not encompass information on ϕ_1 and ψ_{10} .

The three helical states were defined by common values $\chi_n = \chi_{ref}$ for all n (Figure 2). The values $\sigma\chi_{ref}$ were set to $\sigma\chi_\pi = -130^\circ$ (refs 114–117 suggest $\sigma\chi_\pi = -117^\circ, -119^\circ, -127^\circ$, and -131° , respectively), $\sigma\chi_\alpha = -105^\circ$ (refs 118, 118+119 and 118+120+121 suggest $\sigma\chi_\alpha = -103^\circ, -105^\circ$, and -111° , respectively), and $\sigma\chi_{3_{10}} = -75^\circ$ (refs 118+121, 118+119 and 118 suggest $-75^\circ, -78^\circ$, and -89° , respectively). In principle, the above reference values are not sufficient to define ideal regular helices, which are only associated with corresponding χ_n values in the upper right quadrant of the map in Figure 2 (shown in bold). However, the imposition of an additional restraint on the difference between the dihedral angles ϕ_{n+1} and ψ_n (which does not enter into the definition of the reduced conformational subspace) turned out to be unnecessary. In all simulations, these differences never left the range -90° to $+90^\circ$ (for any residue).

Two different forms of biasing potential were considered, labeled P₁ and P₂, the corresponding B&S-LEUS parameters being reported in Tables 1 and 2. The biasing potential P₁ is defined by three spheres centered at the ideal π , α , and 3_{10} conformations, respectively, and connected by a unique (thin) line. The biasing potential P₂ is defined by a unique (thick) line passing through (and extending slightly beyond) the three ideal conformations.

3.3. The “Mother” of All D-Hexopyranoses. A solvated artificial D-hexopyranose was used as another application of the B&S-LEUS scheme to the determination of the relative free energies of the 32 D-hexopyranose stereoisomers (All, Alt, Glc, Man, Gul, Ido, Gal, or Tal; Figure 3), anomers (α or β), and chair conformers (⁴C₁ or ¹C₄), based on a single simulation. This artificial compound, termed here the “mother” of all D-hexopyranoses, consists of a D-hexopyranose where the harmonic improper-dihedral potentials normally controlling the stereochemistry of the hydroxyl groups at carbon atoms C₁, C₂, C₃, and C₄ have been changed to a potential form allowing for the interconversion between the two stereoisomers with a finite energy barrier. As a result, a MD simulation of this artificial compound would in principle provide a good reference ensemble for the perturbative evaluation of the relative free energies of the 32 isomers, by reweighting to the corresponding 16 physical ensembles, considering the two chair forms separately. In practice, however, the epimerization, anomerization, and chair-inversion processes are slow and are enhanced here using the B&S-LEUS approach. The artificial D-hexopyranose was simulated in the presence of $N_{H_2O} = 1200$ water molecules (see section 3.4 for simulation details). The seven-dimensional reduced mixed (conformational and alchemical) subspace was defined by the variables $Q_n = \sigma^{-1}\xi_n$ ($n = 1, \dots, 4$), where ξ_n is the improper-dihedral angle defining the stereochemistry at carbon atom C_n (C₁–O₅–C₂–O₁, C₂–C₁–O₂–C₃, C₃–C₂–C₄–O₃, and C₄–C₅–C₃–O₄), along with $Q_n = \sigma^{-1}\alpha_{n-3}$ ($n = 4, \dots, 7$), where α_1 , α_2 , and α_3 are the

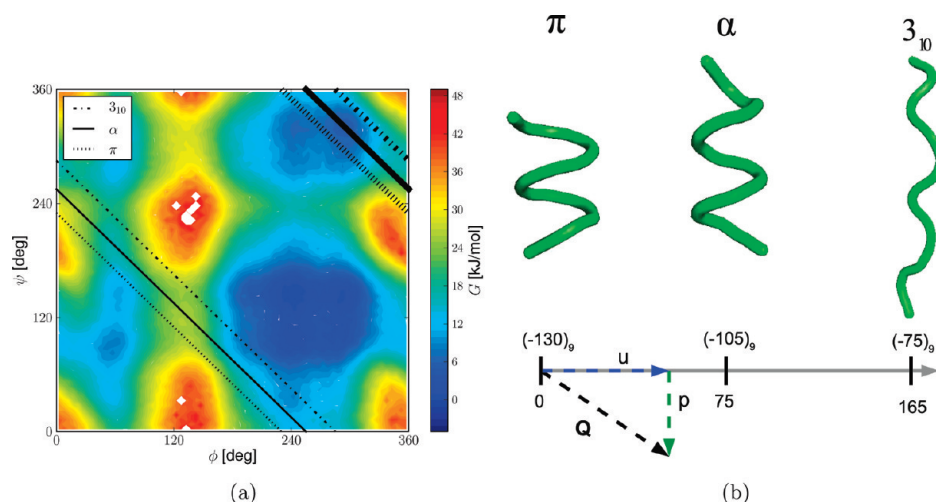


Figure 2. Definition of the reduced conformational subspace and relevant conformational states for the polyalanine decapeptide test system (section 3.2). (a) Definition of ideal conformations for the π , α , and 3_{10} helices, illustrated on a two-dimensional single-linkage basis considering the free-energy surface of a blocked alanine mono-peptide ($\sigma Q_n = \chi_n = \phi_{n+1} + \psi_n$ for $n = 1, \dots, 9$; the line segments defining appropriate values for regular helices, upper right quadrant, are shown in bold). Note that the map is drawn considering a $[0^\circ, 360^\circ]$ dihedral-angle range rather than the more usual $[-180^\circ, 180^\circ]$ range. (b) Representative structures (backbone trace) for the ideal conformations of the three types of helices, and illustration of the longitudinal (u) and transverse (p) distances corresponding to a point Q in the nine-dimensional conformational subspace relative to line \mathcal{L}_1 of biasing potential P_1 (Table 1). The representative points for the ideal helical conformations are $(-130)_9$, $(-105)_9$, and $(-75)_9$ for the π , α , and 3_{10} helices, respectively, where $(Q)_9$ represents a nine-dimensional vector with identical components Q . The corresponding longitudinal distances along line \mathcal{L}_1 are 0, 75 $(-9 \cdot 25^2)^{1/2}$, and 165 $(-9 \cdot 55^2)^{1/2}$, respectively.

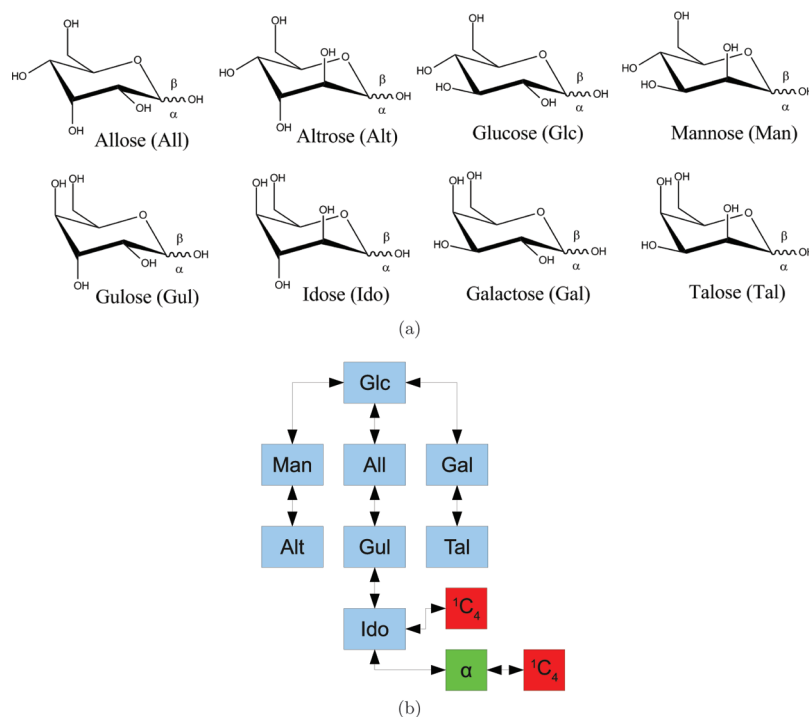


Figure 3. Structures of the hexopyranoses of the D-series and illustration of the maximum-spanning tree employed for the artificial hexopyranose system in protocol H_1 (section 3.3). (a) Structures and naming of the 8 D-hexopyranoses, represented in a 4C_1 chair conformation. (b) Illustration of the maximum-spanning tree employed in protocol H_1 . In this tree, the 4C_1 - β -hexopyranoses are connected by 7 lines (drawn between the blue boxes). Each 4C_1 - β -hexopyranose is further connected to its α -anomer leading to 8 additional lines (schematized by the connection to the green box), and each of the 4C_1 - α - and 4C_1 - β -hexopyranose conformers is connected to the corresponding 1C_4 conformer leading to 16 additional lines (schematized by the connection to the red boxes).

three out-of-plane dihedral angles used to define the ring conformation according to Pickett and Strauss¹²² (values of the improper-dihedral angles $C_4-O_5-C_2-C_1$, $O_5-C_2-C_4-C_3$, and $C_2-C_4-O_5-C_5$, respectively, decreased by 180°).

The 32 states corresponding to the different stereoisomers, anomers, and chair conformers were defined by all possible combinations of the ideal values $\sigma Q_n = \pm 35^\circ$ ($n = 1, \dots, 7$) with the additional constraint $\sigma Q_5 = \sigma Q_6 = \sigma Q_7$ (ideal chair

Table 3. Representation of the 32 D-Hexopyranose Stereoisomers (Figure 3), Anomers (α or β), and Chair Conformers (4C_1 or 1C_4) in Terms of Integer Codes i and Corresponding Bit Strings b_n ^a

i	b_n	isomer	i	b_n	isomer
0	00000	4C_1 - α -idose	16	10000	4C_1 - β -idose
1	00001	1C_4 - α -idose	17	10001	1C_4 - β -idose
2	00010	4C_1 - α -altrose	18	10010	4C_1 - β -altrose
3	00011	1C_4 - α -altrose	19	10011	1C_4 - β -altrose
4	00100	4C_1 - α -talose	20	10100	4C_1 - β -talose
5	00101	1C_4 - α -talose	21	10101	1C_4 - β -talose
6	00110	4C_1 - α -mannose	22	10110	4C_1 - β -mannose
7	00111	1C_4 - α -mannose	23	10111	1C_4 - β -mannose
8	01000	4C_1 - α -gulose	24	11000	4C_1 - β -gulose
9	01001	1C_4 - α -gulose	25	11001	1C_4 - β -gulose
10	01010	4C_1 - α -allose	26	11010	4C_1 - β -allose
11	01011	1C_4 - α -allose	27	11011	1C_4 - β -allose
12	01100	4C_1 - α -galactose	28	11100	4C_1 - β -galactose
13	01101	1C_4 - α -galactose	29	11101	1C_4 - β -galactose
14	01110	4C_1 - α -glucose	30	11110	4C_1 - β -glucose
15	01111	1C_4 - α -glucose	31	11111	1C_4 - β -glucose

^a The coding principle is described in section 3.3.

conformations; 4C_1 , -35° ; 1C_4 , $+35^\circ$). For simplicity, these conformations can be encoded into an integer index i ($i = 1, \dots, 32$) defined by a string of five bits b_n ($n = 1, \dots, 5$, ordered from the highest-weight to the lowest-weight bit), so that $Q_n = 35(2b_n - 1)$ for $n = 1, \dots, 4$ and $Q_5 = Q_6 = Q_7 = 35(2b_5 - 1)$. Thus, for instance, the ideal conformation $i = 17$ (bit string 10001) corresponds to $\xi_1 = +35^\circ$, $\xi_2 = \xi_3 = \xi_4 = -35^\circ$, and $\alpha_1 = \alpha_2 = \alpha_3 = +35^\circ$, that is, to 1C_4 - β -Ido. For the ease of reference, the 32 correspondences between integer indices i and hexopyranose isomers are provided in Table 3.

Two different forms of biasing potential were considered, labeled H₁ and H₂, the corresponding parameters being reported in Tables 1 and 2. Both biasing potentials rely on 32 conformational spheres centered at the corresponding ideal values. In protocol H₁, the 32 spheres are connected by a maximum-spanning tree of 31 lines, illustrated in Figure 3b. Note that this tree is only one possible choice among many others. In protocol H₂ the 32 spheres are connected by a redundant tree of 80 lines motivated by the use of the Manhattan (or Taxicab) metric,¹²³ that is, each state is connected with the five states differing from it by the value of a single bit in terms of binary representation.

3.4. Simulation Details. All MD simulations were carried out using a modified version of the GROMOS05 program¹²⁴ together with the 53A6 force field³⁰ (peptides) and the 56A_{CARBO} force field⁹⁷ (hexopyranose). The systems considered consisted of 1 solute molecule and N_{H_2O} simple point charges¹²⁵ (SPC) water molecules, where N_{H_2O} was 1300, 3300, and 1200 for the blocked alanine mono-peptide, polyalanine decapeptide, and artificial hexopyranose, respectively.

Newton's equations of motion were integrated using the leapfrog algorithm^{126,127} with a 2 fs time step. The SHAKE procedure¹²⁸ was applied to constrain all bond lengths as well as the full rigidity of the solvent molecules with a relative geometric tolerance of 10^{-4} . The simulations were performed under periodic boundary conditions based on a cubic computational box and in the isothermal–isobaric (NPT) ensemble at 298.15 K and 1 bar. The temperature

was maintained by weak coupling of the solute and solvent degrees of freedom (jointly) to a heat bath¹²⁹ with a relaxation time of 0.1 ps. The pressure was maintained by weak coupling of the atomic coordinates and box dimensions to a pressure bath¹²⁹ (isotropic coordinate scaling, group-based virial) with a relaxation time of 0.5 ps and an isothermal compressibility of 0.4575×10^{-3} (kJ mol⁻¹ nm⁻³)⁻¹ as appropriate for water.⁸⁸ The center of mass motion was removed every time step. Nonbonded interactions were handled using a twin-range cutoff scheme,^{2,130} with short- and long-range cutoff distances of 0.8 and 1.4 nm, respectively, and update frequency of 5 time steps for the short-range pairlist and intermediate-range interactions. The mean effect of the omitted electrostatic interactions beyond the long-range cutoff distance was approximately reintroduced using a reaction-field correction,¹³¹ based on a relative dielectric permittivity of 61 as appropriate for the SPC water model.⁴⁰ Configurations (along with U_{bias} in the sampling phase) were written to file every 0.02 ps (peptide) or 0.2 ps (hexopyranose) for subsequent analysis. The systems were preequilibrated using 500 ps MD simulation.

As described in section 3.3, the simulation of the “mother” of all D-hexopyranoses required a modification of the functional form for the terms controlling the stereochemistry at carbon atoms C₁, C₂, C₃, and C₄. In the GROMOS force field,^{88,89,124} the chirality of a center is determined by an improper-dihedral angle energy term of the form

$$V_\xi(\xi) = \frac{1}{2}k_\xi(\xi - \xi_0)^2 \quad (28)$$

where two opposite values of ξ_0 characterize the two stereoisomers. In the simulations of the “mother” of all D-hexopyranoses, this interaction term was modified on the basis of the EDS principle⁷⁸ to

$$V'_\xi(\xi) = -\frac{1}{\beta s'} \ln \left\{ \exp[-\beta s' \frac{1}{2}k_\xi(\xi - |\xi_{01}|)^2] + \exp[-\beta s' \frac{1}{2}k_\xi(\xi + |\xi_{02}|)^2] \right\} \quad (29)$$

where s' is a barrier-smoothing parameter. These choices, along with the setting $\beta s' = 0.1$, led to smooth anomerization and epimerization transitions, with barriers on the order of 50–100 kJ mol⁻¹. The thermodynamic quantities appropriate for the physical ensemble considering one specific stereoisomer are then calculated on the basis of the configurations generated during the sampling phase of a B&S-LEUS simulation, via a reweighting procedure that now depends on the sum of the biasing potential \mathcal{U}_{bias} and of the difference $V'_\xi - V_\xi$ for this specific stereoisomer.^{15,97} Note that this specific application of the EDS principle is functionally entirely distinct from its use in the unification of the single-object biasing potentials described in section 2.4.

4. Results

4.1. Blocked Alanine Mono-peptide. The results of the 10 simulations (A₁–A₁₀; Tables 1 and 2) involving a (blocked) alanine mono-peptide (section 3.1) are displayed in Figure 4 in the form of Ramachandran free-energy maps.

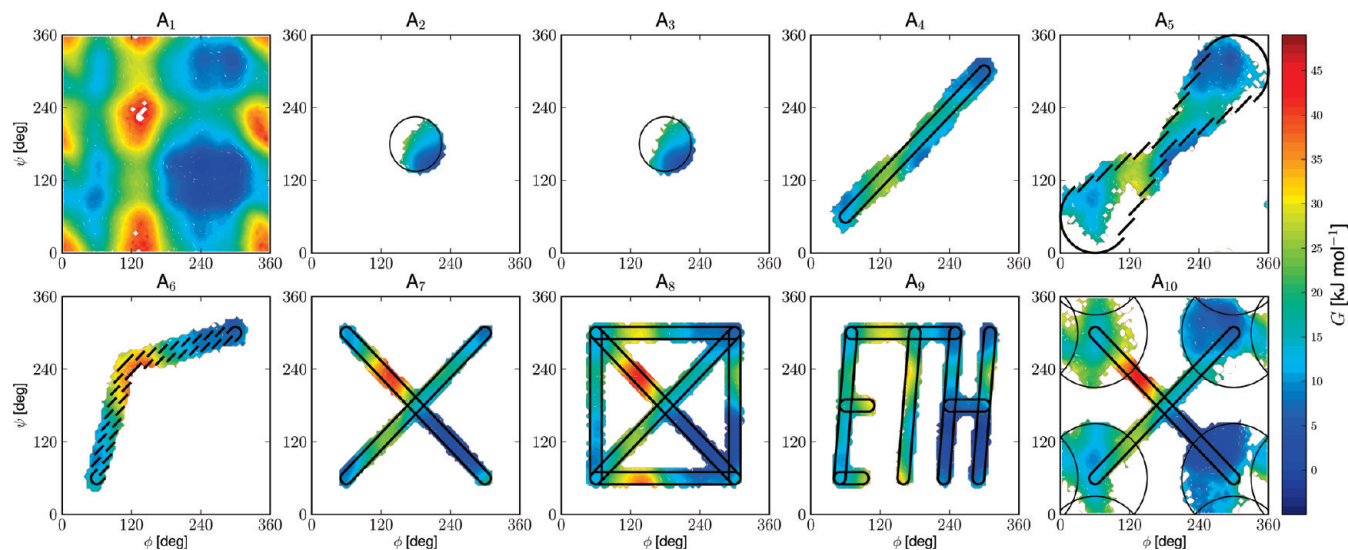


Figure 4. Free-energy maps obtained from the different B&S-LEUS simulations (along with one plain LEUS simulation) of a blocked alanine mono-peptide in water (section 3.1). The relevant conformational subspace is defined by $\sigma(Q_1, Q_2) = (\phi, \psi)$, and the maps represent the calculated free energy. The characteristics of the conformational objects (spheres or/and lines) defining the B&S-LEUS biasing potentials and the parameters of the associated protocol are provided in Tables 1 and 2, respectively, for the different simulations (A₁–A₁₀). The maps are calculated by reweighting of the biased probability distribution using a grid spacing of 5°, and anchored by fitting of the covered area onto the free-energy map for simulation A₁, itself anchored to zero at the location of its global minimum. Note that the maps are drawn considering [0°, 360°] dihedral-angle ranges rather than the more usual [−180°, 180°] ranges.

Simulations A₂–A₁₀ are B&S-LEUS simulations, while simulation A₁ is a reference plain LEUS simulation from ref 37 showing the entire free-energy map.

The results illustrate, in the simple context of a reduced conformational subspace of low dimensionality, the versatility of the B&S-LEUS scheme in the design of different patterns of active subspaces (here, active surfaces) in terms of spheres (here, disks) and lines. They also underline that the B&S-LEUS sampling is nearly homogeneous radially but not necessarily directionally within the spheres (e.g., A₂, A₃, and A₁₀), as well as longitudinally but not necessarily transversally within the lines (e.g., A₅). Note that the latter inhomogeneities will be accentuated when considering problems of higher dimensionalities. The most remarkable feature in Figure 4 is that all maps are closely resemblant to each other within the active subspace. The achievement of the B&S-LEUS procedure is to permit a flexible (problem-adapted) definition of this subspace, guarantee an appropriate sampling within it, and exclude its exterior from the sampling.

Simulation A₁₀ can be considered as a two-dimensional prototype of what the B&S-LEUS approach is meant to achieve in higher-dimensionality problems, namely the connection of a number of conformational states (here, four) by means of lines (here, two) ensuring a sufficient number of transitions between these states. Further details concerning this simulation are shown in Figure 5.

The number of visits N_v to different conformational points (squares of edge 5°) observed during the sampling phase of this simulation is displayed in Figure 5a (on a logarithmic scale). Projections along the line directions or along the sphere radii (data not shown) evidence, as expected, essentially homogeneous distributions of N_v , although the two-

dimensional distribution itself shows a significant extent of inhomogeneity within the active subspace, especially inside the spheres. The biasing potential $\mathcal{U}_{\text{bias}}$ obtained at the end of the LE build-up phase (and used during the sampling phase) is shown in Figure 5b. As expected, due to the action of the restraining potentials, any sampling outside the active subspace is essentially precluded. Within the active subspace, the memory-based component compensates for the physical bias in the free-energy surface (i.e., it is higher in regions where the free energy is lower). Although one might consider using the negative of $\mathcal{U}_{\text{bias}}$ as an approximate estimate for the free energy G (eq 25), the biased map $G + \mathcal{U}_{\text{bias}}$ (Figure 5d) shows that the corresponding error can be very large (up to about ± 10 kJ mol^{−1} for specific conformational points).

Simulations A₅ and A₆ illustrate two variants of lines, the line with longitudinally dependent width and the displaced line (section 2.3). Simulations A₂ and A₃ illustrate two variants of biasing potentials within a sphere, either radially homogeneous or with a radial distribution proportional to the Jacobian factor (here, to the distance to the sphere center) using a nonunit distribution-alteration function (eq 23 instead of eq 22). Further details concerning the latter two simulations are shown in Figure 6.

The numbers of visits N_v per conformational point (squares of edge 5°) for the two simulations are shown in Figure 6a and b (on a logarithmic scale), and the corresponding radial projections are shown in Figure 6c and d. As expected, the nonunit distribution-alteration function biases the sampling of the sphere toward its periphery and changes the homogeneous radial distribution to an approximately linear one. However, in view of the low dimensionality and long

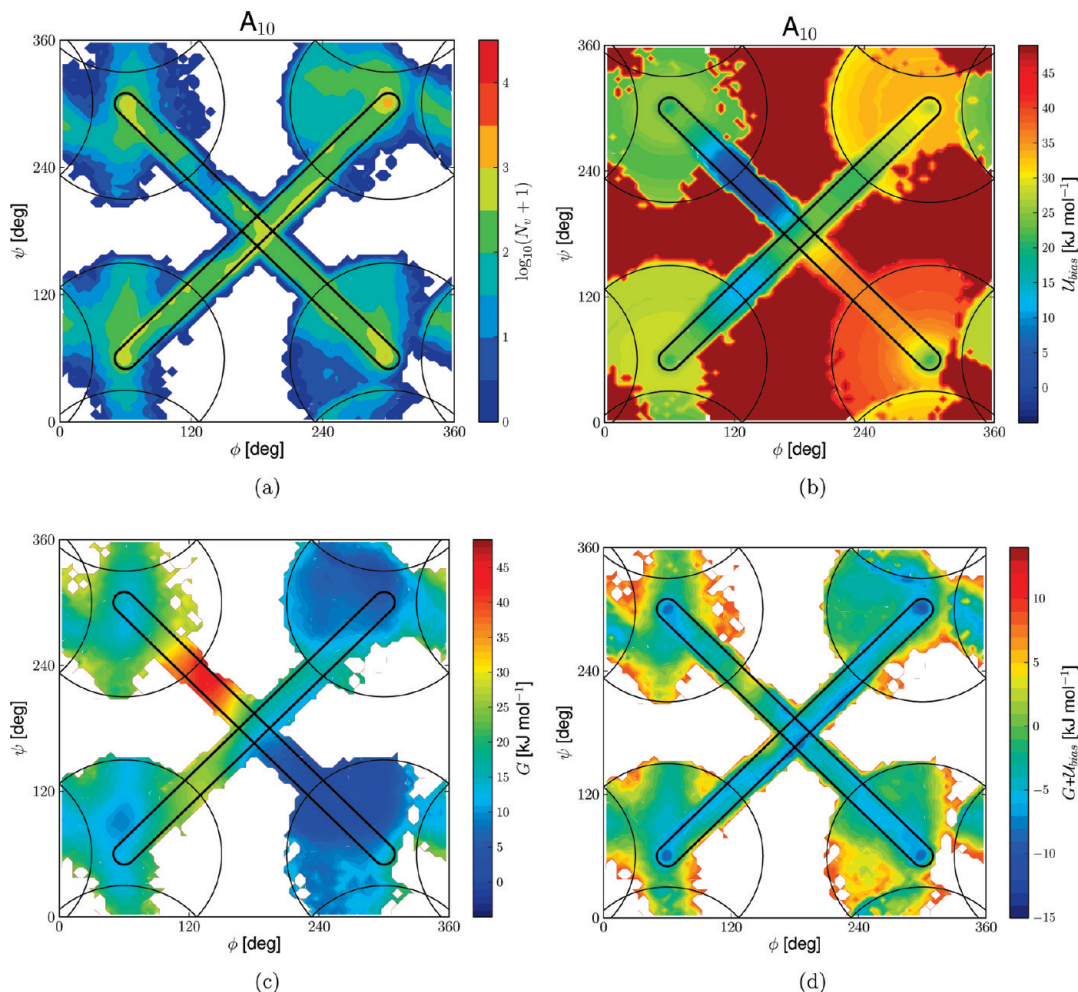


Figure 5. Illustrative details of the B&S-LEUS simulation of a blocked alanine mono-peptide in water using the biasing potential A_{10} , see legend of Figure 4. (a) Number of visits N_v (displayed as $\log_{10}(N_v + 1)$) to a given conformational point during the sampling phase. (b) Optimized biasing potential $\mathcal{U}_{\text{bias}}$ (all points that were not sampled, or where $\mathcal{U}_{\text{bias}} \geq 50$ kJ mol $^{-1}$, are shown in red). (c) Calculated free-energy map G . Biased free-energy map $G + \mathcal{U}_{\text{bias}}$. The maps and numbers of visits are calculated using a grid spacing of 5° . The map of panel (c) is identical to that displayed in Figure 4. The sum of N_v for panel (a) is equal to 250 000 (5 ns sampling phase, coordinates written every 0.02 ps).

sampling time, this change does not significantly alter the final free-energy map (Figure 4).

4.2. Polyalanine Decapeptide. The results of the two B&S-LEUS simulations (P_1 and P_2 ; Tables 1 and 2) involving an (unblocked) polyalanine decapeptide (section 3.2) are displayed in Figures 7 and 8. The biasing potential P_1 is defined by three spheres centered at the ideal π , α , and 3_{10} conformations, respectively, and connected by an unique (thin) line. The biasing potential P_2 is defined by a unique (thick) line passing through (and extending slightly beyond) the three ideal conformations.

The time evolutions and distributions of the longitudinal distance u and transverse distance p (with reference to line \mathcal{L}_1 of protocol P_1) are displayed in Figure 7a and b for simulations P_1 and P_2 , respectively.

The longitudinal distance u accounts for one degree of freedom of the reduced conformational subspace. In both simulations, it varies over a range (-26 to 262 in P_1 , -122 to 226 in P_2) appropriate to cover the three types of helices (π , 0 ; α , 75 ; 3_{10} , 165). However, the number of “sweeps” across this interval remains limited, and the corresponding (biased) distributions $P(u)$ are not homogeneous, suggesting

that a sampling phase longer than 2×50 ns (and possibly also a longer build-up phase) might be desirable. Note that a homogeneous distribution is expected in simulation P_2 upon full convergence, but not necessarily in simulation P_1 due to the presence of the three spheres overlapping with the line. In both simulations, the regions corresponding to the π and α helices are found to be significantly more sampled than that corresponding to the 3_{10} helix (the more limited sampling of the 3_{10} helix region is most pronounced in simulation P_2).

The transverse distance p accounts for the remaining eight degrees of freedom of the reduced conformational subspace. This distance displays much lower fluctuations along the trajectories, with relatively narrow (biased) distributions $P(p)$ centered at about 30 and 100 for simulations P_1 and P_2 , respectively. A lower value for P_1 is expected due to the inclusion of the spheres, which enforce a radial biasing relative to three reference points characterized by $p = 0$. The average values of 30 and 100 correspond to root-mean-square deviations of about 10° and 35° in terms of the individual dihedral-angle differences χ_n .

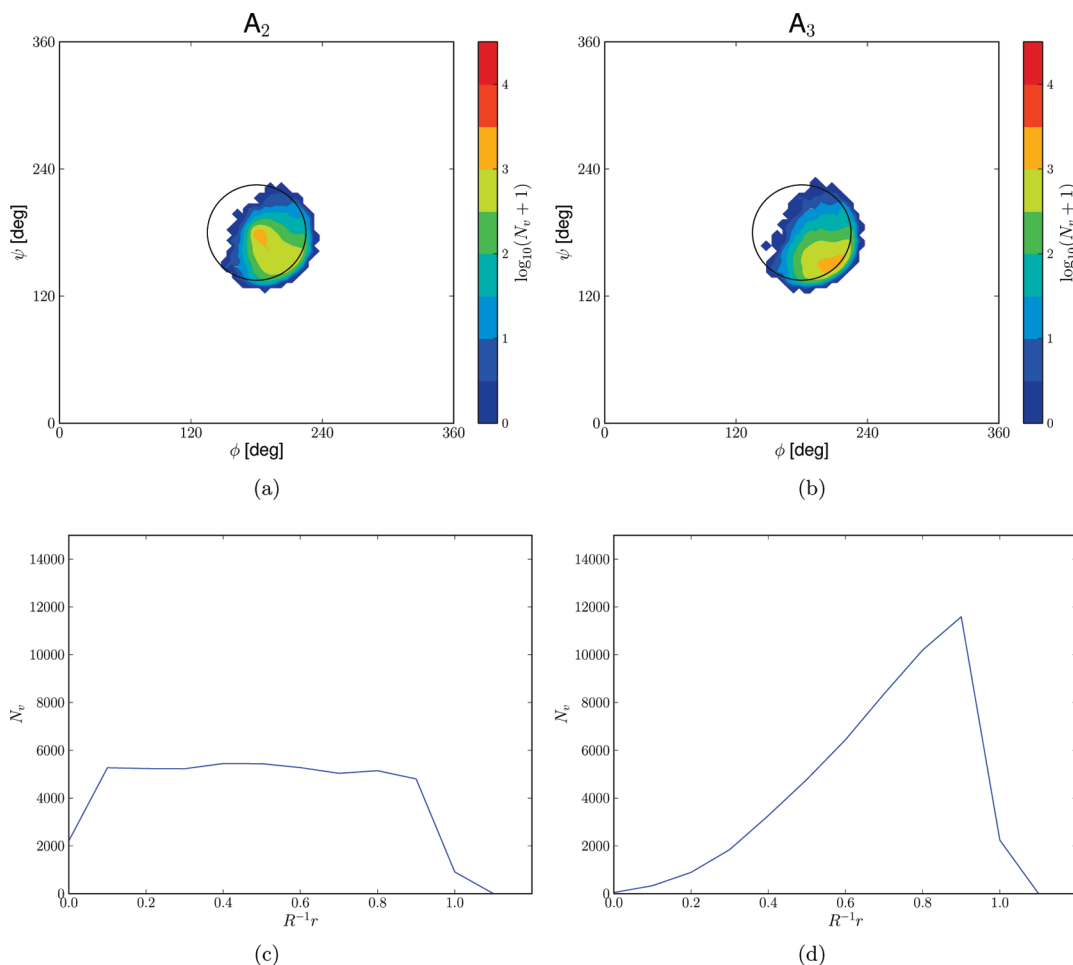


Figure 6. Illustrative details of the B&S-LEUS simulations of a blocked alanine mono-peptide in water using the biasing potentials A_2 and A_3 , see legend of Figure 4. (a,b) Number of visits N_v (displayed as $\log_{10}(N_v + 1)$) to a given conformational point during the sampling phases of simulations A_2 (a) and A_3 (b). (c,d) Number of visits N_v as a function of the ratio of the distance from the sphere center to the sphere radius during the sampling phases of simulations A_2 (c) and A_3 (d). The numbers of visits are calculated using a grid spacing of 5° . The sum of N_v for panels (a) and (b) is equal to 250 000 (5 ns sampling phase, coordinates written every 0.02 ps).

The time evolutions of the nine dihedral-angle differences χ_n defining the reduced conformational subspace ($\sigma^{-1}Q_n = \chi_n$) and of the local DSSP secondary-structure assignment¹³² at the eight central residues m ($m = 2, \dots, 9$, where the residue m corresponds to ϕ_m and ψ_m and is thus comprised between χ_{m-1} and χ_m) are displayed in Figure 7c and d for simulations P_1 and P_2 , respectively.

The individual dihedral-angle differences χ_n evidence large and highly correlated fluctuations. This correlation is expected, considering that the biasing potential has been designed to drive the sampling along a “diagonal” line in the nine-dimensional reduced conformational subspace. The magnitude of the fluctuations is higher and the extent of correlation lower for the terminal χ_1 and χ_9 values. The probable reasons for this effect are: (i) a less pronounced dependence of the free energy on the orientation of the (more flexible) terminal residues; and (ii) a correlation of these dihedral-angle differences with the dihedral angles ϕ_1 and ψ_{10} that are not included in the biasing scheme.

The local DSSP secondary-structure assignment reveals that in a significant fraction of the configurations where the longitudinal distance u as well as the individual χ_n values are in the ideal π or α helical regions, a π or α helix is

actually formed, despite the nonzero transverse distance p . In contrast, only a very limited fraction of the configurations where u and the χ_n values are in the ideal 3_{10} region are actually recognized as 3_{10} helices, due to the transverse distance p . In other words, the nonzero transverse distance p predominantly accounts for a conformational variability of the simulated configurations in the neighborhood of ideal π or α structures, while it predominantly represents a conformational discrepancy of the simulated configurations with respect to an ideal 3_{10} helix structure.

The number of visits N_v to a given conformational point (rectangles of edges $\Delta u = 7$ and $\Delta p = 2.4$), as well as the free-energy maps in the subspace of the longitudinal and transverse (or longitudinal only) distances u and p , are displayed in Figure 8a,c,e,g and b,d,h,f for simulations P_1 and P_2 , respectively. These two-dimensional (or one-dimensional) maps provide simplified variants (projections) of the corresponding nine-dimensional maps in the full reduced conformational subspace, which obviously cannot be easily visualized.

For simulation P_1 , the number of visits (Figure 8a) reveals an extensive coverage of the two-dimensional subspace (within the bounds -50 to $+250$ for u and 0 to $+125$ for p).

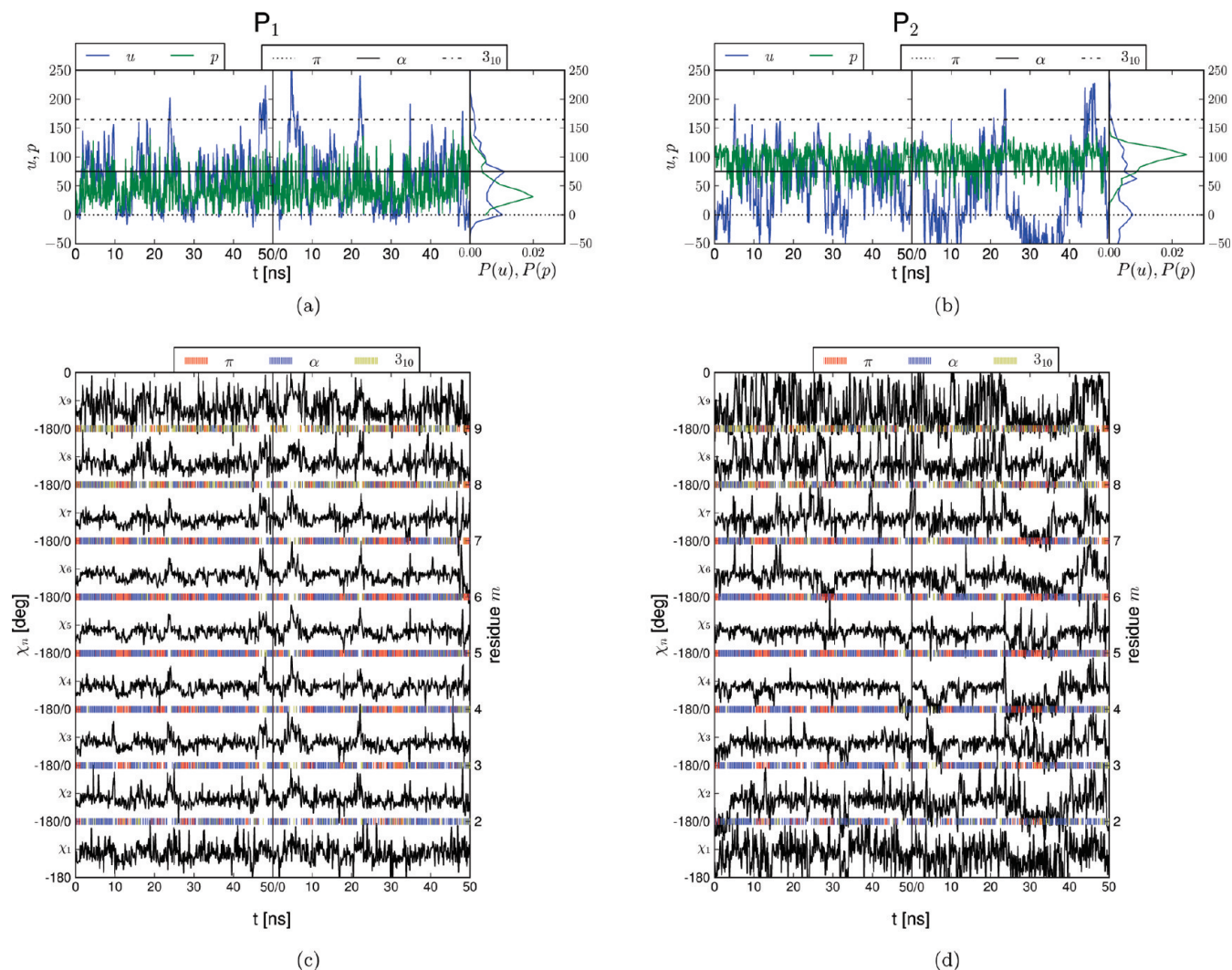


Figure 7. Time series of the longitudinal and transverse distances along a line connecting the relevant helical states, as well as of the conformational coordinates and local helical patterns, for the two B&S-LEUS simulations of a polyaniline decapeptide in water (section 3.2). The relevant conformational subspace is defined by $\sigma Q_n = \chi_n = \phi_{n+1} + \psi_n$ ($n = 1, \dots, 9$), see Figure 2. (a,b) Time series (and corresponding normalized distributions in the biased ensemble) of the longitudinal (u) and transverse (p) distances, referring to the line \mathcal{L}_1 of protocol P_1 (Table 1 and Figure 2), for the sampling phases of simulations P_1 (a) and P_2 (b). (c,d) Time series of the conformational coordinates χ_n , and of the local helical pattern (assigned according to DSSP¹³²), for the sampling phases of simulations P_1 (c) and P_2 (d). Note that the results correspond to the concatenated sampling phases of two independent 50 ns simulations.

As discussed above (Figure 7a), this coverage is not entirely homogeneous along u and shows a predominance of visited conformations around $p = 30$. For simulation P_2 (Figure 7b), however, a significant portion of the subspace in the neighborhood of the ideal 3_{10} helix conformation ($u = 165$, $p = 0$) is not visited.

The reason for this difference becomes obvious when comparing the corresponding free-energy maps (Figure 8c and d). As is clearly visible from the results of simulation P_1 , the close neighborhood of the 3_{10} helix corresponds to a region of very high free energy. In simulation P_1 , the sampling of this neighborhood is enforced by the inclusion of a corresponding conformational sphere. In simulation P_2 , the sampling of this region is not enforced, because the longitudinal sampling homogeneity within the line is compatible with high p values when u is close to 165. Besides the above difference in the extent of coverage, the maps issued from the two independent simulations are remarkably

similar inside the regions they both cover. This suggests a sufficient level of convergence in the two simulations, despite the residual inhomogeneity in $P(u)$ (see above).

The configurations recognized by the DSSP algorithm as π , α , and 3_{10} helices (for 7 out of 8 residues at least) are shown superimposed on the maps in Figure 8e and f for simulations P_1 and P_2 , respectively. A higher number of configurations are recognized as regular helices for simulation P_1 as compared to simulation P_2 , due to the lower average value of p (Figure 7a vs b), a consequence of the use of the three spheres in P_1 . Note again that configurations recognized as π or α helix can still present relatively high p values (variability relative to corresponding ideal conformations), while very few configurations are recognized as 3_{10} helix (almost none for simulation P_2).

Finally, the free energy and average value \bar{p} of p are shown as a function of u in Figure 8g and h for simulations P_1 and P_2 , respectively. Note that \bar{p} is not the straight average of p

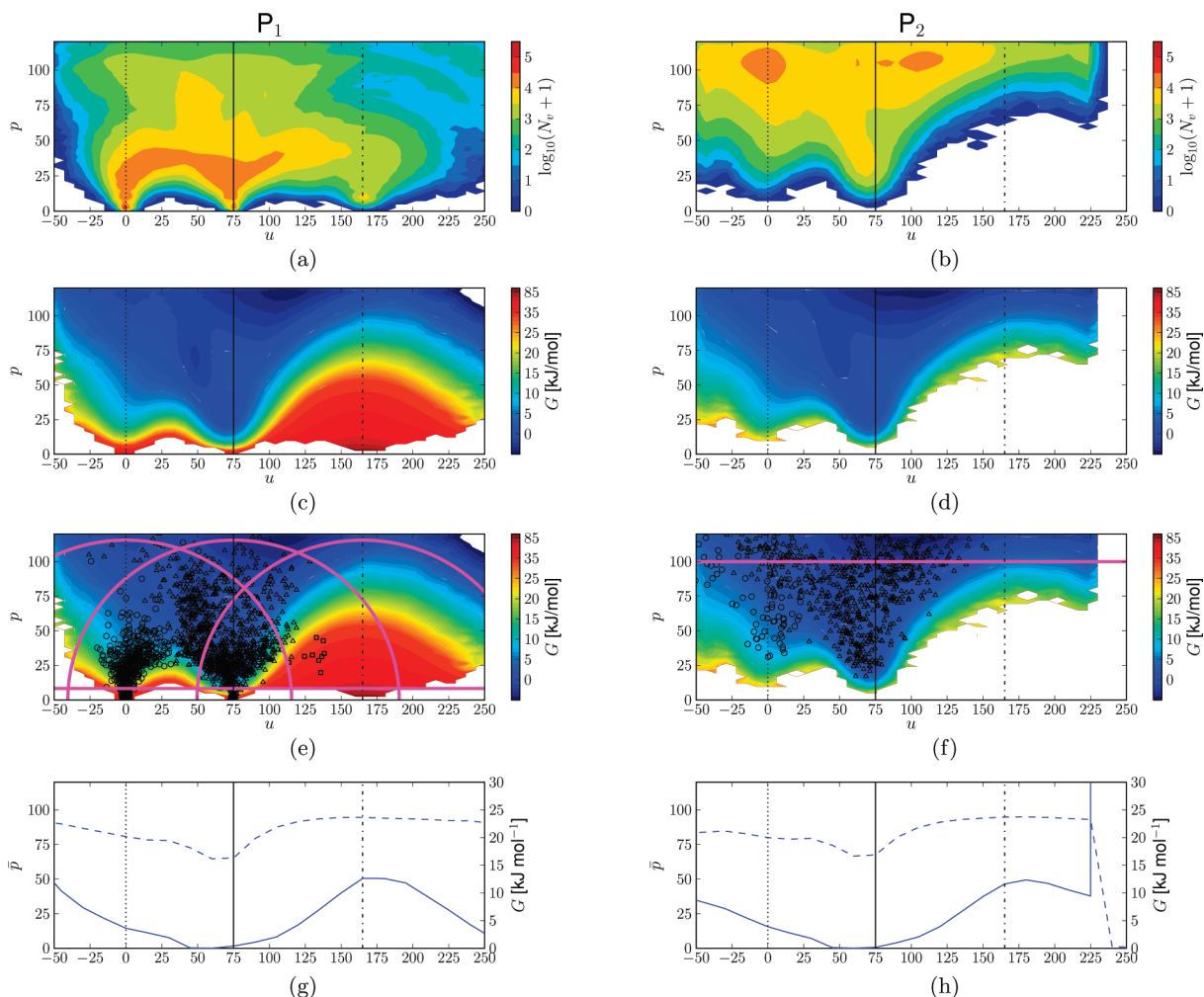


Figure 8. Number of visits N_v and free-energy maps (longitudinal-transverse and purely longitudinal projections) corresponding to a line connecting the relevant helical states for the two B&S-LEUS simulations of a polyaniline decapeptide in water (section 3.2). The relevant conformational subspace is defined by $\sigma Q_n = \chi_n = \phi_{n+1} + \psi_n$ ($n = 1, \dots, 9$), see Figure 2. The longitudinal (u) and transverse (p) components refer to the line \mathcal{L}_1 of protocol P_1 (Table 1 and Figure 2). (a,b) Number of visits N_v during the sampling phase of the simulations P_1 (a) or P_2 (b), displayed on a logarithmic scale. (c,d,e,f) Free energy $G(u, p)$ in a longitudinal-transverse projection for the simulations P_1 (c,e) or P_2 (d,f). (g,h) Free energy $G(u)$ (solid line) and average transverse distance $\bar{p}(u)$ (dashed line; reweighted to the physical ensemble) in a purely longitudinal projection for the simulations P_1 (g) or P_2 (h). The maps and number of visits are calculated using a grid spacing $\Delta u = 7$ and $\Delta p = 2.4$. The sum of N_v for panels (a) and (b) is equal to 5×10^6 (100 ns sampling phase, coordinates written every 0.02 ps). Panels (c) and (d) are identical to panels (e) and (f), respectively. However, in the latter, the spheres and lines involved in the B&S-LEUS potentials (lines representing their peripheries, based on the corresponding radii and widths) are shown as solid lines, as well as representative points of conformations assigned by DSSP¹³² as π (circles), α (triangles), and 3_{10} (squares) helices (at least seven linkages assigned to the corresponding helical conformation out of eight; Figure 7). Note that the color coding of the free energy is not linear for high free energies.

along the biased simulations (Figure 7a and b), due to reweighting to the physical ensemble. The two curves are closely similar in the two simulations (over the range they jointly sample). They indicate relative free energies of the order of 4, 0, and 12 kJ mol^{-1} for the π , α , and 3_{10} helical forms of a polyaniline decapeptide in water based on the GROMOS 53A6 force field.³⁰ They also suggest the possibility of a smooth (barrier-free) interconversion pathway along the cooperative coordinate u . Of course, this implies neither that this path is the lowest free-energy path, nor that most interconversion trajectories follow this path. Finally, the average transverse distance \bar{p} is nearly constant along u (about 85, involving root-mean-square single-residue angular deviations of the order of 30°), except close to the α -helix (corresponding values of about 70 and 25° , respectively).

As a final note, the noticeable (but still limited) discrepancies between dihedral-angle and DSSP descriptions for helical structures suggest that it might be interesting to repeat the present simulations based on a conformational subspace defined by hydrogen-bonding rather than backbone dihedral-angle coordinates.

4.3. The “Mother” of All D-Hexopyranoses. The results of the two B&S-LEUS simulations (H_1 and H_2 ; Tables 1 and 2) involving the artificial D-hexopyranose (section 3.3) are displayed in Figures 9–11. Both biasing potentials rely on 32 conformational spheres centered at the corresponding ideal values. In protocol H_1 , the 32 spheres are connected by a maximum-spanning tree of 31 lines, illustrated in Figure 3b. In protocol H_2 , the 32 spheres are connected by a redundant tree of 80 lines motivated by the use of the

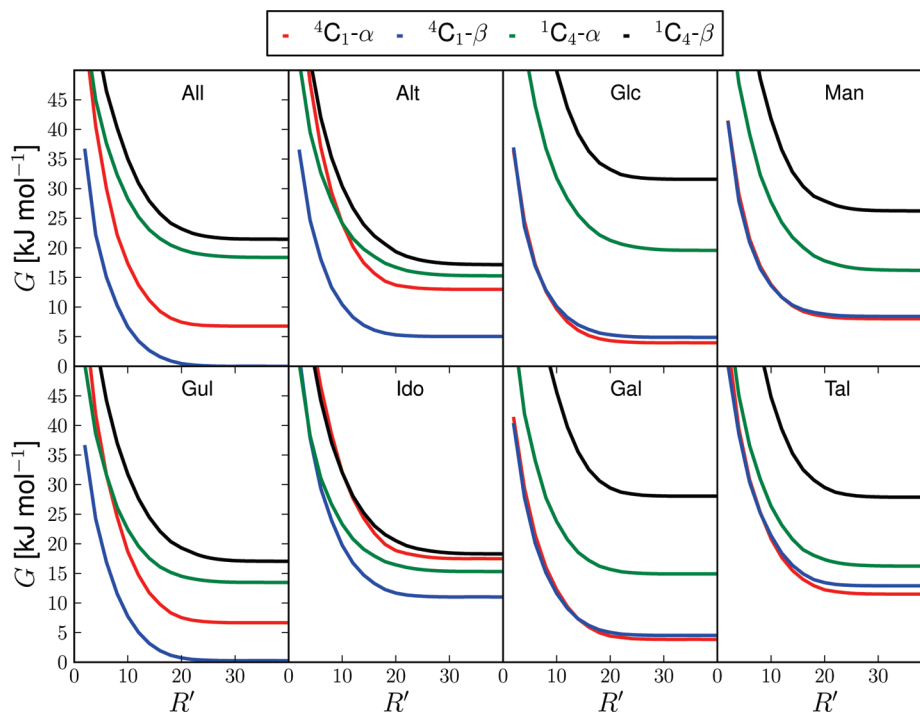


Figure 9. Relative free energies of the 32 different isomers as a function of the sphere radius used for the assignment of the conformations to states in the B&S-LEUS simulations of the “mother” of all D-hexopyranoses in water (section 3.3) using protocol H₂ (Tables 1 and 2). The assignment to states is based on seven-dimensional spheres of radius R' centered at the 32 ideal D-hexopyranose stereoisomers (Figure 3), anomers, and chair conformers. The free energies of the different states are calculated on the basis of the (reweighted) population of a state relative to the total (reweighted) population of the entire reduced conformational space, and offset so as to bring the free energy of ${}^4C_{1-\beta}$ -All to zero for large R' .

Manhattan (or Taxicab) metric,¹²³ that is, each state is connected with the five states differing from it by the value of a single bit in terms of binary representation.

The relative free energies calculated for the 32 isomers based on simulation H₂ are shown in Figure 9 as a function of the radius R' used to assign sampled conformations to states. The corresponding curves for simulation H₁ are qualitatively similar (data not shown). As discussed in section 2.7, the volumes used for the state assignment need not be identical to those used in the definition of the biasing potential. In the present case, both types of volumes are spheres centered at the 32 ideal conformations, with radii $R = 50$ for the former ones (Table 1) and R' for the latter ones. Note that the free-energy values in this figure are normalized relative to the total (reweighted) population of the reduced conformational subspace (and offset so as to bring the lowest value of ${}^4C_{1-\beta}$ -All to zero at large R').

For small R' , the free energies of the states diverge to $+\infty$ as the fractional populations they encompass become infinitesimally small. Note that in this limit, the corresponding differences remain in principle finite, but become increasingly inaccurate (and ultimately diverge) as a consequence of finite sampling. For sufficiently large R' , the free energies converge to well-defined values, indicating that a further extension of the sphere radii only adds high free-energy regions, which no longer contribute significantly to the free energy of the states, that is, the corresponding local free-energy basins have been fully encompassed. Note that the ideal conformations may not exactly correspond to the free-energy minima of these local basins. However, this is expected to have a minor

influence on the results as long as R' is chosen sufficiently large (and given that the basins corresponding to different states remain well separated). Based on the results of Figure 9, spheres of radii $R' = 30$ were chosen for all subsequent state assignments.

The results of the simulations of the artificial hexopyranose based on protocols H₁ and H₂ are shown in Figure 10a and b, respectively. Both simulations have visited the entire set of 32 states during their sampling phases. The fraction of intermediate configurations that could not be assigned to any state is about 30%. This number is actually remarkably low, that is, only about one-third of the computational effort has been invested in ensuring a sufficient number of transitions between the states via statistically irrelevant regions. Note that this fraction is not higher for simulation H₂ as compared to simulation H₁, although the latter involves more lines (80 vs 31). On the other hand, simulation H₂ achieved a much higher extent of homogeneity in terms of state populations within the biased ensemble and in terms of the numbers of transitions between these states. As a result, the relative free energies evaluated for the different states based on protocol H₂ are expected to be more precise (as supported by the lower error bars) and more accurate (within the employed force field and simulation methodology).

The convergence and accuracy properties of the two protocols are compared in Figure 11, in terms of the chair-inversion free-energy change $\Delta G_{C_1 \rightarrow C_4}$. Figure 11a shows the root-mean-square deviation (rmsd) between estimates based on the sampling phase up to time t and corresponding (well-converged) estimates from the 16 independent LEUS

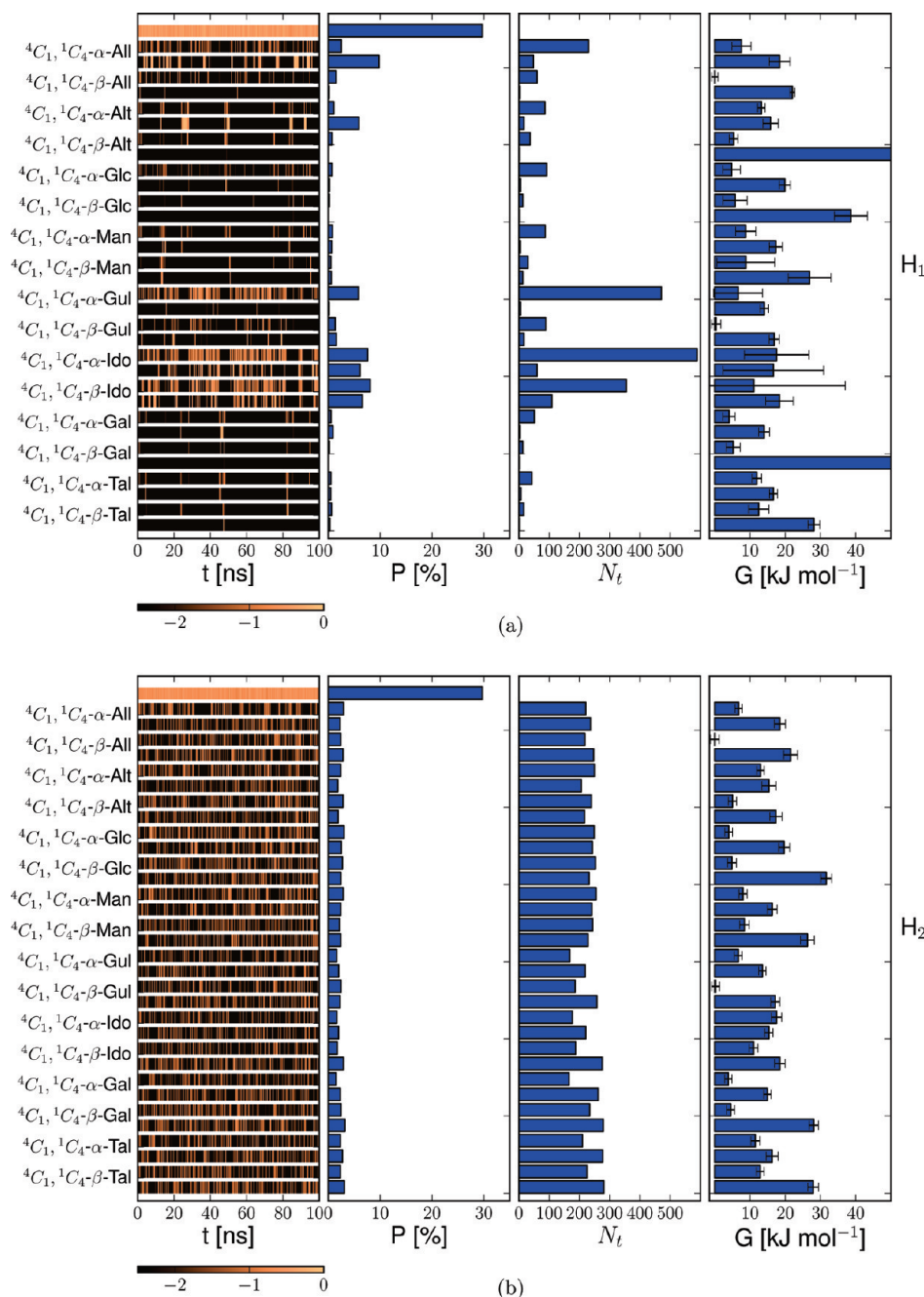


Figure 10. Results of the B&S-LEUS simulations of the “mother” of all D-hexopyranoses in water (section 3.3). (a) Results of the simulations H₁. (b) Results of the simulations H₂. The following quantities are displayed (from left to right), with reference to the 32 ideal D-hexopyranose stereoisomers (Figure 3), anomers, and chair conformers: the fraction (on a decimal logarithmic scale) of successive 40 ps intervals along the sampling phase spent visiting a given state, the relative probability P of a given state in the biased ensemble during the sampling phase, the number of transitions N_t observed to a given state from any other state during the sampling phase, and the relative free energies G of the states after reweighting. The visit fractions and free energies rely on a state assignment using seven-dimensional spheres of radius $R' = 30$ centered at the corresponding ideal conformations.

simulations reported in ref 97. Because the free energy of states that have not yet (or only poorly) been sampled at time t is formally infinite, the rmsd calculation only includes the N_p changes for which the number of visits to both conformers up to time t is at least 100. The detailed comparison in terms of values corresponding to the entire sampling phase is shown in Figure 11b.

Protocol H₁ shows a somewhat erratic convergence to a final rmsd value of 1.0 kJ mol⁻¹, with three identified

outliers (β -Alt, β -Glc, and β -Gal; absolute deviations of 43.7, 6.4, and 30.9 kJ mol⁻¹, respectively; not included in the rmsd), corresponding to the states that have been poorly visited during the biased simulation (Figure 10a). In contrast, protocol H₂ shows a smooth convergence to a final rmsd value of 0.7 kJ mol⁻¹ (all states included). Here, the deviation is homogeneously spread across the 16 free-energy changes, and within the statistical error bars in nearly all cases.

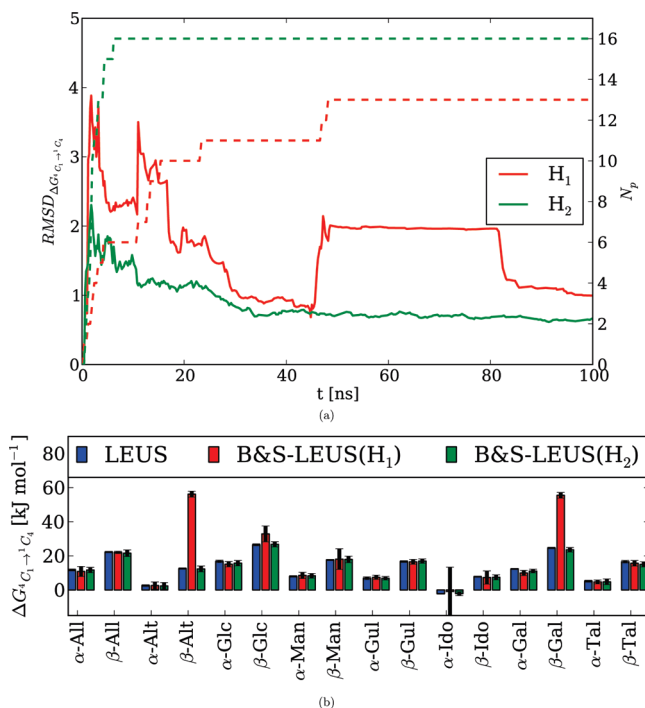


Figure 11. Comparison of the relative free energies of D-hexopyranose chair conformers in water calculated using a single B&S-LEUS simulations of the “mother” of all D-hexopyranoses (section 3.3) or using 16 plain LEUS simulations. The B&S-LEUS results correspond to simulations H_1 and H_2 (Tables 1 and 2). The plain LEUS results are from ref 97. (a) Root-mean-square deviation (rmsd; solid lines) between the free-energy difference $\Delta G_{C_1 \rightarrow C_4}$ as predicted by the B&S-LEUS simulation (protocols H_1 or H_2) considering a sampling time t and the corresponding values obtained using the 16 plain LEUS simulations. Only pairs of states that have been both visited more than 100 times are included in the rmsd calculation. The corresponding number of pairs N_p is also shown (dashed lines). (b) Free-energy difference $\Delta G_{C_1 \rightarrow C_4}$ between the 4C_1 and 1C_4 ring conformers evaluated using the 16 plain LEUS simulations or the two B&S LEUS simulations.

The above results suggest that the use of a redundant tree (H_2) rather than a maximum-spanning tree (H_1) leads to improved convergence in the present case. Note, however, that an extension of the simulation length (sampling and, possibly, build-up) beyond 100 ns could almost certainly lead to converged results also using protocol H_1 . Conversely, accurate results could be obtained using protocol H_2 with (sampling and, possibly, build-up) times shorter than 100 ns. Finally, it is worth emphasizing that the B&S-LEUS scheme has led to accurate estimates for the 16 relevant free-energy differences in a single simulation involving two phases of 100 ns each. In contrast, the plain LEUS estimates⁹⁷ required 16 independent simulations, each involving two phases of 3 and 40 ns each (total 688 ns) for a comparable precision. For comparison, corresponding estimates using plain US might well require 160 simulations (assuming 10 windows per simulation). In addition, the B&S-LEUS simulations also provide the relative free energies between anomers and epimers, the calculation of which would require 15 additional LEUS or sets of US simulations.

5. Conclusion

The present work is concerned with the development of a new method, ball-and-stick local elevation umbrella sampling (B&S-LEUS), to enhance the sampling in computer simulations of (bio)molecular systems. This approach enables in particular the calculation of conformational free-energy differences between states even in situations where the definition of these states relies on a conformational subspace involving more than a few degrees of freedom. The basic principle is to associate spheres (“balls”) to all relevant states and define a set of lines (“sticks”) that connect them, the union of all of these objects defining an active conformational subspace. A biasing potential involving confinement restraints (to restrict the sampling within the active subspace) and a memory-based term (to enforce a nearly homogeneous sampling of the active subspace, radially within spheres and longitudinally within lines) is then constructed using the local elevation (LE) procedure and applied in a subsequent umbrella sampling (US) simulation. The performance of the B&S-LEUS approach was tested here in the context of three illustrative examples: (i) the restriction of the sampling to arbitrary areas of the Ramachandran map of a (blocked) monopeptide in water; (ii) the evaluation of the relative free energies of three helical forms of a solvated polyanaline decapeptide; and (iii) the calculation of the relative free energies of the 32 isomers, anomers, and chair conformers of aqueous D-hexopyranoses. This new approach is appealing for the following three main reasons.

First, the B&S-LEUS method is generally applicable to virtually any type of reduced subspace definition and state properties to be evaluated. The degrees of freedom involved in the reduced subspace definition can be of conformational (internal coordinate of the physical system), alchemical (extended-state λ variable of λ -dynamics^{91–96}), or even thermodynamic (e.g., temperature or pressure) nature, or a any mixture of the three. The corresponding state properties can be the free energy (present work), but also any other thermodynamic (e.g., enthalpy, entropy, heat capacity, or volume) or structural (ensemble average of a given instantaneous observable) quantity.

Second, the B&S-LEUS method is a problem-oriented (engineering) scheme, which makes it suitable for the direct translation of precise scientific questions. Typical questions such as “what are the relative thermodynamic properties of a set of distinct conformational states of a given macromolecule”, “what are the relative solvation properties of a set of molecules in a given solvent”, “what are the relative binding properties of a set of molecules to a given receptor”, or “how are the simulated properties of a system altered upon changing a given set of force-field parameters” can in principle directly be translated into the language of “balls” and “sticks”, resulting in a question-specific scheme (i.e., a scheme requiring the minimal possible computational effort to answer only this specific question). Note that in the conformational context, the translation of what is meant experimentally by the word “state” (e.g., ensemble of conformations characterized by specific spectroscopic or functional properties) into a choice of reduced coordinates and conformational volumes may still hide a significant extent of complexity and ambiguity (see below).

Third, the B&S-LEUS method is in large parts “automated”. The human component resides in the choice of the reduced space, in the definition of the various “balls” and “sticks”, in the specification of a number of protocol parameters (prominently, the build-up basis force constant and force-reduction factor, the build-up and sampling durations, and, possibly, a state-assignment scheme), and in the convergence assessment and data analysis. As a simple example, the present simulations of the “mother” of all D-hexopyranoses returned 31 (converged) free-energy differences in one single simulation (2×100 ns). A corresponding evaluation of the 31 stepwise changes via thermodynamic integration or umbrella sampling, assuming 20 λ points or windows per change, would have required about 600 simulations (probably for about the same total sampling duration or more), and a considerable amount of human effort (adjustment of the staging, equilibration and sampling protocols, baby-sitting of the computer jobs, combination of the data from the individual simulations, and consideration of case-to-case problems).

Fourth, the B&S-LEUS method addresses simultaneously the searching efficiency problem (i.e., how to avoid the continuous revisiting of previously discovered configurations), the statistical efficiency problem (i.e., how to avoid compromising the optimal Boltzmann-weighted sampling of plain MD as much as possible), the transition problem (i.e., how to ensure a high number of transitions between the relevant states), and the computational efficiency problem (i.e., how to avoid an excessive computational overhead and memory requirement). As a result, for a given design of the problem geometry as defined by the various “balls” and “sticks”, this scheme is likely to provide a close-to-optimal compromise in terms of sampling efficiency, that is, a close-to maximal achievable accuracy for a given user-specified problem and a given amount of computer time.

The four above considerations suggest that the B&S-LEUS approach has a very high potential for practical applications in the area of molecular simulation. However, although it efficiently addresses the computational part of the problem, both the setup (definition of the reduced conformational subspace and active region thereof) and the post-processing (state assignment) may involve intricate issues related to the very definition of the concept of state. For example, depending on the adopted perspective, a state might be defined either as a single free-energy basin (thermodynamic definition), as a collection of configurations with low interconversion barriers (kinetic definition), as a collection of configurations with low mutual coordinate deviations (structural definition), or as a conformational region associated with given spectroscopic or functional properties (experimental definition). The choice of one type of definition or another is ultimately in the hands of the chemist, but it should be realized that different definitions may lead to very different interpretations of the simulation results.

Acknowledgment. We would like to thank Zhixiong Lin for his help during the setup of the polyalanine decapeptide simulations. Financial support from the Swiss National Science Foundation (Grant NF200021-121895) is also gratefully acknowledged.

References

- (1) Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids*; Oxford University Press: New York, 1987.
- (2) van Gunsteren, W. F.; Berendsen, H. J. C. *Angew. Chem., Int. Ed. Engl.* **1990**, *29*, 992–1023.
- (3) van Gunsteren, W. F.; Bakowies, D.; Baron, R.; Chandrasekhar, I.; Christen, M.; Daura, X.; Gee, P.; Geerke, D. P.; Glättli, A.; Hünenberger, P. H.; Kastenholz, M. A.; Oostenbrink, C.; Schenk, M.; Trzesniak, D.; van der Vegt, N. F. A.; Yu, H. B. *Angew. Chem., Int. Ed.* **2006**, *45*, 4064–4092.
- (4) Berendsen, H. J. C. *Simulating the Physical World*; Cambridge University Press: Cambridge, U.K., 2007.
- (5) Rick, S. W.; Stuart, S. J. *Rev. Comput. Chem.* **2002**, *18*, 89–146.
- (6) Yu, H.; van Gunsteren, W. F. *Comput. Phys. Commun.* **2005**, *172*, 69–85.
- (7) Stern, H. A.; Berne, B. J. *J. Chem. Phys.* **2001**, *115*, 7622–7628.
- (8) Geerke, D. P.; Lubber, S.; Marti, K. H.; van Gunsteren, W. F. *J. Comput. Chem.* **2008**, *30*, 514–523.
- (9) Hünenberger, P. H.; van Gunsteren, W. F. In *Computer Simulation of Biomolecular Systems, Theoretical and Experimental Applications*; van Gunsteren, W. F., Weiner, P. K., Wilkinson, A. J., Eds.; Kluwer/Escom Science Publishers: Dordrecht, The Netherlands, 1997; pp 3–82.
- (10) Kastenholz, M.; Hünenberger, P. H. *J. Phys. Chem. B* **2004**, *108*, 774–788.
- (11) Reif, M. M.; Kräutler, V.; Kastenholz, M. A.; Daura, X.; Hünenberger, P. H. *J. Phys. Chem. B* **2009**, *113*, 3112–3128.
- (12) van Gunsteren, W. F.; Huber, T.; Torda, A. E. *AIP Conf. Proc.* **1995**, *330*, 253–268.
- (13) Berne, B. J.; Straub, J. E. *Curr. Opin. Struct. Biol.* **1997**, *7*, 181–189.
- (14) Christen, M.; van Gunsteren, W. F. *J. Comput. Chem.* **2008**, *29*, 157–166.
- (15) Hansen, H. S.; Hünenberger, P. H. *J. Comput. Chem.* **2010**, *31*, 1–23.
- (16) Beveridge, D. L.; DiCapua, F. M. *Annu. Rev. Biophys. Biophys. Chem.* **1989**, *18*, 431–492.
- (17) Straatsma, T. P.; McCammon, J. A. *Annu. Rev. Phys. Chem.* **1992**, *43*, 407–435.
- (18) King, P. M. In *Computer Simulation of Biomolecular Systems, Theoretical and Experimental Applications*; van Gunsteren, W. F., Weiner, P. K., Wilkinson, A. J., Eds.; ESCOM Science Publishers: B.V., Leiden, The Netherlands, 1993; Vol. 26, pp 7–314.
- (19) van Gunsteren, W. F.; Beutler, T. C.; Fraternali, F.; King, P. M.; Mark, A. E.; Smith, P. E. In *Computer Simulation of Biomolecular Systems, Theoretical and Experimental Applications*; van Gunsteren, W. F., Weiner, P. K., Wilkinson, A. J., Eds.; ESCOM Science Publishers: B.V., Leiden, The Netherlands, 1993; Vol. 31, pp 5–367.
- (20) Kollman, P. *Chem. Rev.* **1993**, *93*, 2395–2417.
- (21) Straatsma, T. P. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; VCH Publishers Inc.: New York, 1996; Vol. 8, pp 1–127.

- (22) Chipot, C.; Pearlman, D. A. *Mol. Simul.* **2002**, *28*, 1–12.
- (23) Rodinger, T.; Pomès, R. *Curr. Opin. Struct. Biol.* **2005**, *15*, 164–170.
- (24) Gilson, M. K.; Given, J. A.; Bush, B. L.; McCammon, J. A. *Biophys. J.* **1997**, *72*, 1047–1069.
- (25) Wang, J.; Deng, Y.; Roux, B. *Biophys. J.* **2006**, *91*, 2798–2814.
- (26) Scheraga, H. A.; Khalili, M.; Liwo, A. *Annu. Rev. Phys. Chem.* **2007**, *58*, 57–83.
- (27) van Gunsteren, W. F.; Gattin, Z. In *Foldamers: Structure, Properties and Applications*; Hecht, S., Huc, I., Eds.; Wiley: Weinheim, Germany, 2007.
- (28) Kastenholz, M. A.; Schwartz, T. U.; Hünenberger, P. H. *Biophys. J.* **2006**, *91*, 2976–2990.
- (29) Villa, A.; Mark, A. E. *J. Comput. Chem.* **2002**, *23*, 548–553.
- (30) Oostenbrink, C.; Villa, A.; Mark, A. E.; van Gunsteren, W. F. *J. Comput. Chem.* **2004**, *25*, 1656–1676.
- (31) Geerke, D. P.; van Gunsteren, W. F. *ChemPhysChem* **2006**, *7*, 671–678.
- (32) Oostenbrink, C.; van Gunsteren, W. F. *Proteins: Struct., Funct., Genet.* **2004**, *54*, 234–246.
- (33) Christ, C. D.; van Gunsteren, W. F. *J. Comput. Chem.* **2009**, *30*, 1664–1679.
- (34) Hünenberger, P. H.; Granwehr, J. K.; Aebischer, J.-N.; Ghoneim, N.; Haselbach, E.; van Gunsteren, W. F. *J. Am. Chem. Soc.* **1997**, *119*, 7533–7544.
- (35) Torrie, G. M.; Valleau, J. P. *J. Comput. Phys.* **1977**, *23*, 187–199.
- (36) Valleau, J. P.; Torrie, G. M. In *Modern Theoretical Chemistry*; Berne, B. J., Ed.; Plenum Press: New York, 1977; Vol. 16, pp 9–194.
- (37) Hansen, H. S.; Daura, X.; Hünenberger, P. H. *J. Chem. Theory Comput.*, companion manuscript; DOI: 10.1021/ct1003059.
- (38) Beutler, T. C.; van Gunsteren, W. F. *J. Chem. Phys.* **1994**, *100*, 1492–1497.
- (39) Kumar, S.; Rosenberg, J. M.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A. *J. Comput. Chem.* **1995**, *16*, 1339–1350.
- (40) Heinz, T. N.; van Gunsteren, W. F.; Hünenberger, P. H. *J. Chem. Phys.* **2001**, *115*, 1125–1136.
- (41) Piccinini, E.; Ceccarelli, M.; Affinito, F.; Brunetti, R.; Jacoboni, C. *J. Chem. Theory Comput.* **2008**, *4*, 173–183.
- (42) Paine, G. H.; Scheraga, H. A. *Biopolymers* **1985**, *24*, 1391–1436.
- (43) Mezei, M. *J. Comput. Phys.* **1987**, *68*, 237–248.
- (44) Hoof, R. W. W.; van Eijck, B. P.; Kroon, J. *J. Chem. Phys.* **1992**, *97*, 6690–6694.
- (45) Friedman, R. A.; Mezei, M. *J. Chem. Phys.* **1995**, *102*, 419–426.
- (46) Bartels, C.; Karplus, M. *J. Comput. Chem.* **1997**, *18*, 1450–1462.
- (47) Wang, J.; Gu, Y.; Liu, H. *J. Chem. Phys.* **2006**, *125*, 094907/1–094907/9.
- (48) Babin, V.; Roland, C.; Darden, T. A.; Sagui, C. *J. Chem. Phys.* **2006**, *125*, 204909/1–204909/9.
- (49) Marsili, S.; Barducci, A.; Chelli, R.; Procacci, P.; Schettino, V. *J. Phys. Chem. B* **2006**, *110*, 14011–14013.
- (50) Lelièvre, T.; Rousset, M.; Stoltz, J. *Chem. Phys.* **2007**, *126*, 134111/1–134111/8.
- (51) van der Vaart, A.; Karplus, M. *J. Chem. Phys.* **2007**, *126*, 164106/1–164106/17.
- (52) Babin, V.; Roland, C.; Sagui, C. *J. Chem. Phys.* **2008**, *128*, 134101/1–134101/7.
- (53) Barnett, C. B.; Naidoo, K. *J. Mol. Phys.* **2009**, *107*, 1243–1250.
- (54) Laio, A.; Parrinello, M. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 12562–12566.
- (55) Laio, A.; Rodriguez-Fortea, A.; Gervasio, F. L.; Ceccarelli, M.; Parrinello, M. *J. Phys. Chem. B* **2005**, *109*, 6714–6721.
- (56) Darve, E.; Rodriguez-Gomez, D.; Pohorille, A. *J. Chem. Phys.* **2008**, *128*, 144120/1–144120/13.
- (57) Kirkwood, J. G. *J. Chem. Phys.* **1935**, *3*, 300–313.
- (58) Straatsma, T. P.; McCammon, J. A. *J. Chem. Phys.* **1991**, *95*, 1175–1188.
- (59) Zwanzig, R. W. *J. Chem. Phys.* **1954**, *22*, 1420–1426.
- (60) Tobias, D. J.; Brooks, C. L., III. *Chem. Phys. Lett.* **1987**, *142*, 472–476.
- (61) Mitchell, M. J.; McCammon, J. A. *J. Comput. Chem.* **1991**, *12*, 271–275.
- (62) Simonson, T. *Mol. Phys.* **1993**, *80*, 441–447.
- (63) Beutler, T. C.; Mark, A. E.; van Schaik, R.; Gerber, P. R.; van Gunsteren, W. F. *Chem. Phys. Lett.* **1994**, *222*, 529–539.
- (64) Fixman, M. *Proc. Natl. Acad. Sci. U.S.A.* **1974**, *71*, 3050–3053.
- (65) Helfand, E. *J. Chem. Phys.* **1979**, *71*, 5000–5007.
- (66) Boresch, S.; Karplus, M. *J. Chem. Phys.* **1996**, *105*, 5145–5154.
- (67) Straatsma, T. P.; Zacharias, M.; McCammon, J. A. *Chem. Phys. Lett.* **1992**, *196*, 297–302.
- (68) den Otter, W. K.; Briels, W. J. *J. Chem. Phys.* **1998**, *109*, 4139–4146.
- (69) Hermans, J.; Shankar, S. *Isr. J. Chem.* **1986**, *27*, 225–227.
- (70) Oostenbrink, C.; van Lipzig, M. M. H.; van Gunsteren, W. F. In *Comprehensive Medicinal Chemistry II Computer-Assisted Drug Design*; Taylor, J. B., Triggle, D. J., Eds.; Elsevier: Amsterdam, The Netherlands, 2007; Vol. 65, pp 1–668.
- (71) Pitera, J. W. *Curr. Opin. Drug Discovery Dev.* **2009**, *12*, 388–396.
- (72) Smith, P. E.; van Gunsteren, W. F. *J. Chem. Phys.* **1994**, *100*, 577–585.
- (73) Schäfer, H.; van Gunsteren, W. F.; Mark, A. E. *J. Comput. Chem.* **1999**, *20*, 1604–1617.
- (74) Pitera, J. W.; van Gunsteren, W. F. *J. Phys. Chem.* **2001**, *105*, 11264–11274.
- (75) Oostenbrink, C.; van Gunsteren, W. F. *J. Comput. Chem.* **2003**, *24*, 1730–1739.
- (76) Oostenbrink, C.; van Gunsteren, W. F. *Chem.-Eur. J.* **2005**, *11*, 4340–4348.

- (77) Oostenbrink, C.; van Gunsteren, W. F. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6750–6754.
- (78) Christ, C. D.; van Gunsteren, W. F. *J. Chem. Phys.* **2007**, *126*, 184110/1–184110/10.
- (79) Christ, C. D.; van Gunsteren, W. F. *J. Chem. Phys.* **2008**, *128*, 174112/1–174112/12.
- (80) Christ, C. D.; van Gunsteren, W. F. *J. Chem. Theory Comput.* **2009**, *5*, 276–286.
- (81) Crippen, G. M.; Scheraga, H. A. *Chemistry* **1969**, *64*, 42–49.
- (82) Levy, A. V.; Montalvo, A. *SIAM J. Sci. Stat. Comput.* **1985**, *6*, 15–29.
- (83) Glover, F. *ORSA J. Comput.* **1989**, *1*, 190–206.
- (84) Huber, T.; Torda, A. E.; van Gunsteren, W. F. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 695–708.
- (85) Grubmüller, H. *Phys. Rev. E* **1995**, *52*, 2893–2906.
- (86) Engkvist, O.; Karlström, G. *Chem. Phys.* **1996**, *213*, 63–76.
- (87) Fukunishi, Y.; Mikami, Y.; Nakamura, H. *J. Phys. Chem. B* **2003**, *107*, 13201–13210.
- (88) van Gunsteren, W. F.; Billeter, S. R.; Eising, A. A.; Hünenberger, P. H.; Krüger, P.; Mark, A. E.; Scott, W. R. P.; Tironi, I. G. *Biomolecular Simulation: The GROMOS96 manual and user guide*; Verlag der Fachvereine: Zürich, Switzerland, 1996.
- (89) Scott, W. R. P.; Hünenberger, P. H.; Tironi, I. G.; Mark, A. E.; Billeter, S. R.; Fennen, J.; Torda, A. E.; Huber, T.; Krüger, P.; van Gunsteren, W. F. *J. Phys. Chem. A* **1999**, *103*, 3596–3607.
- (90) Perić-Hassler, L.; Hansen, H. S.; Baron, R.; Hünenberger, P. H. *Carbohydr. Res.* **2010**, *345*, 1781–1801.
- (91) Tidor, B. *J. Phys. Chem.* **1993**, *97*, 1069–1073.
- (92) Kong, X.; Brooks, C. L., III. *J. Chem. Phys.* **1996**, *105*, 2414–2423.
- (93) Guo, Z.; Brooks, C. L., III; Kong, X. *J. Phys. Chem. B* **1998**, *102*, 2032–2036.
- (94) Guo, Z.; Brooks, C. L., III. *J. Am. Chem. Soc.* **1998**, *120*, 1920–1921.
- (95) Leitgeb, M.; Schröder, C.; Boresch, S. *J. Chem. Phys.* **2005**, *122*, 084109/1–084109/15.
- (96) Abrams, J. B.; Rosso, L.; Tuckerman, M. E. *J. Chem. Phys.* **2006**, *125*, 074115/1–074115/12.
- (97) Hansen, H. S.; Hünenberger, P. H. *J. Comput. Chem.*, in press.
- (98) Carter, E. A.; Ciccotti, G.; Hynes, J. T.; Kapral, R. *Chem. Phys. Lett.* **1989**, *156*, 472–477.
- (99) Depaepe, J.-M.; Ryckaert, J.-P.; Paci, E.; Ciccotti, G. *Mol. Phys.* **1993**, *79*, 515–522.
- (100) Schlitter, J.; Engels, M.; Krüger, P.; Jacoby, E.; Wollmer, A. *Mol. Simul.* **1993**, *10*, 291–308.
- (101) Gō, N.; Scheraga, H. A. *Macromolecules* **1976**, *9*, 535–542.
- (102) Gottlieb, M.; Bird, R. B. *J. Chem. Phys.* **1976**, *65*, 2467–2468.
- (103) Kannan, S.; Zacharias, M. *Proteins: Struct., Funct., Bioinf.* **2007**, *66*, 697–706.
- (104) Xu, C.; Wang, J.; Liu, H. *J. Chem. Theory Comput.* **2008**, *4*, 1348–1359.
- (105) Liu, Z.; Berne, B. J. *J. Chem. Phys.* **1993**, *99*, 6071–6077.
- (106) Barducci, A.; Bussi, G.; Parrinello, M. *Phys. Rev. Lett.* **2008**, *100*, 020603/1–020603/4.
- (107) Henkelman, G.; Jónsson, H. *J. Chem. Phys.* **2000**, *113*, 9978–9985.
- (108) Maragliano, L.; Fischer, A.; Vanden-Eijnden, E.; Ciccotti, G. *J. Chem. Phys.* **2006**, *125*, 024106/1–024106/15.
- (109) Sheppard, D.; Terrel, R.; Henkelman, G. *J. Chem. Phys.* **2008**, *128*, 134106/1–134106/10.
- (110) Vanden-Eijnden, E.; Venturoli, M. *J. Chem. Phys.* **2009**, *130*, 194103/1–194103/17.
- (111) Schlitter, J.; Swegat, W.; Mülders, T. *J. Mol. Model.* **2001**, *7*, 171–177.
- (112) Schlitter, J.; Engels, M.; Krüger, P. *J. Mol. Graphics* **1994**, *12*, 84–89.
- (113) Makino, Y.; Itoh, N. *BMC Struct. Biol.* **2008**, *8*, 46/1–46/15.
- (114) Fodje, M. N.; Al-Karadaghi, S. *Protein Eng.* **2002**, *15*, 353–358.
- (115) Al-Karadaghi, S.; Cedergren-Zeppezauer, E. S.; Hovmöller, S.; Petratos, K.; Terry, H.; Wilson, K. S. *Acta Crystallogr., Sect. D* **1994**, *50*, 793–807.
- (116) Low, B. W.; Grenville-Wells, H. J. *Chemistry* 19539785801.
- (117) Lee, K.-H.; Benson, D. R.; Kuczera, K. *Biochemistry* **2000**, *39*, 13737–13747.
- (118) Barlow, D. J.; Thornton, J. M. *J. Mol. Biol.* 1988201601619.
- (119) Pauling, L. *Proc. Natl. Acad. Sci. U.S.A.* 195137235240.
- (120) Arnott, S.; Wonacott, A. J. *J. Am. Chem. Soc.* **1966**, *88*, 2598–2599.
- (121) Perutz, M. F. *Nature* **1951**, *167*, 1053–1054.
- (122) Pickett, H. M.; Strauss, H. L. *J. Am. Chem. Soc.* **1970**, *92*, 7281–7290.
- (123) Krause, E. F. *Taxicab Geometry*; Dover Publications: New York, 1986.
- (124) Christen, M.; Hünenberger, P. H.; Bakowies, D.; Baron, R.; Bürgi, R.; Geerke, D. P.; Heinz, T. N.; Kastenholz, M. A.; Kräutler, V.; Oostenbrink, C.; Peter, C.; Trzesniak, D.; van Gunsteren, W. F. *J. Comput. Chem.* **2005**, *26*, 1719–1751.
- (125) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Hermans, J. In *Intermolecular Forces*; Pullman, B., Ed.; Reidel: Dordrecht, The Netherlands, 1981; Vol. 33, pp 1–342.
- (126) Feynman, R. P.; Leighton, R. B.; Sands, M. *The Feynman Lectures on Physics*; Addison-Wesley: Boston, MA, 1963.
- (127) Hockney, R. W. *Methods Comput. Phys.* **1970**, *9*, 136–211.
- (128) Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *23*, 327–341.
- (129) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Di Nola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684–3690.
- (130) Berendsen, H. J. C.; van Gunsteren, W. F.; Zwinderman, H. R. J.; Geurtsen, R. G. *Ann. N. Y. Acad. Sci.* 1986482269285.
- (131) Tironi, I. G.; Sperb, R.; Smith, P. E.; van Gunsteren, W. F. *J. Chem. Phys.* **1995**, *102*, 5451–5459.
- (132) Kabsch, W.; Sander, C. *Biopolymers* **1983**, *22*, 2577–2637.

JCTC

Journal of Chemical Theory and Computation

G4-SP, G4(MP2)-SP, G4-sc, and G4(MP2)-sc: Modifications to G4 and G4(MP2) for the Treatment of Medium-Sized Radicals

Bun Chan,^{*,†,‡} Michelle L. Coote,^{†,§} and Leo Radom^{*,†,‡}

*ARC Center of Excellence for Free Radical Chemistry and Biotechnology,
School of Chemistry, University of Sydney, NSW 2006, Australia, and Research School
of Chemistry, Australian National University, ACT 0200, Australia*

Received May 19, 2010

Abstract: The G4-SP and G4(MP2)-SP procedures are introduced, as alternatives to G4 and G4(MP2), to overcome shortcomings associated with the treatment of spin polarization (SP) in large open-shell systems. The new methods employ a converging SP term, to replace the diverging A' treatment used in the G4 and G4(MP2) formulations. The G4-SP and G4(MP2)-SP procedures have mean absolute deviations (MADs) from experimental energies of 3.49 and 4.37 kJ mol⁻¹, respectively, for the G3/05 test set, which are comparable to the MAD values for G4 and G4(MP2) but eliminate the problem of a diverging A' term. For energies involving larger radicals, G4(MP2)-SP performs better than standard G4(MP2). Alternative methods, including G4-5H, G4(MP2)-5H, G4-sc, and G4(MP2)-sc, are also introduced to avoid the problem of an indefinitely increasing SP correction in standard G4 or G4(MP2) for reactions involving larger open-shell systems.

1. Introduction

The Gaussian series (G_n) of composite quantum chemistry methods were formulated by Curtiss, Raghavachari, and Pople et al. for the accurate evaluation of thermochemical properties using a series of relatively inexpensive component calculations.¹ The latest among the G_n series, namely G4,^{1d} has achieved impressive accuracy, with a mean absolute deviation (MAD) from experimental values of 3.48 kJ mol⁻¹ (0.83 kcal mol⁻¹) for the G3/05 test set² of 456 energies. More economical versions of G_n , termed $G_n(\text{MP2})$,³ have also been developed, that have a slightly lower accuracy. For instance, G4(MP2)^{3c} produces an MAD of 4.36 kJ mol⁻¹ (1.04 kcal mol⁻¹) for the G3/05 test set.

An empirical correction term called the “higher level correction” (HLC) is an integral part of the G_n and $G_n(\text{MP2})$

methods and was introduced to compensate for any residual deficiencies of the methods. For G3^{1c} and G3(MP2),^{3b} the HLC for molecular species takes the form $-An_\beta - B(n_\alpha - n_\beta)$, where A and B are parameters derived from empirical fitting to reliable experimental thermochemical data and n_α and n_β are the number of valence α and β electrons. As n_α and n_β are equal for closed-shell species, the HLC is simplified to $-An_\beta$ for such molecules. For G4 and G4(MP2), this component of the HLC for closed-shell species has the same form as that for G3, i.e., $-An_\beta$. For open-shell molecular species, however, a slightly different HLC has been introduced, which takes the form $-A'n_\beta - B(n_\alpha - n_\beta)$. The additional A' parameter was introduced in order to account for deficiencies in the treatment of open-shell systems.

It has recently been pointed out⁴ that the new form of HLC in G4 and G4(MP2) introduces a correction term for many radical reactions, such as radical additions and hydrogen abstractions, which was not present in earlier G_n procedures such as G3. For example, it is easy to show that for a generic radical-addition reaction:

* Corresponding author e-mail: chan_b@chem.usyd.edu.au (B.C.); radom@chem.usyd.edu.au (L.R.).

† ARC Center of Excellence for Free Radical Chemistry and Biotechnology.

‡ University of Sydney.

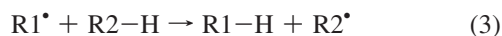
§ Australian National University.



the HLC correction with G4 or G4(MP2) is

$$(A - A')n_{\beta}(X) \quad (2)$$

Similarly, for a hydrogen-abstraction reaction:



the HLC contribution to the reaction is

$$(A - A')[n_{\beta}(R2^{\bullet}) - n_{\beta}(R1^{\bullet})] \quad (4)$$

Equation 2 indicates that for an addition reaction, the G4 HLC contribution is directly proportional to the size of X, the molecule to which R^{\bullet} is to be added. For a hydrogen-abstraction reaction, eq 4 indicates that the correction is proportional to the difference in the size of R1 and R2. Thus, as the size of the species in a radical reaction increases, the standard G4 HLC correction to the reaction energy may grow to a very large number.

This growth in the correction term seems somewhat counterintuitive. For example, in the stabilization energy reaction,



we might expect a substantial cancellation of errors in the calculation of the reaction energy using methods such as CCSD(T) with a large basis set, giving good results without the need for an empirical correction. However, standard G4 would lead to an HLC correction for reaction 5 of $-(18 + 9n)(A - A')$, which increases rapidly with n . For example, for $n = 10$, the HLC correction amounts to $-51.3 \text{ kJ mol}^{-1}$.

In the present study, we attempt to address this issue by examining three ways of reducing or eliminating the dependence of the HLC on the size of the system. First, we examine the effect of not having a separate A' parameter (i.e., as is the case in G3), thus reducing the number of empirical parameters to 5. This gives rise to G4-5H and G4(MP2)-5H (5H for 5 HLC parameters). Second, we examine whether the type of spin-contamination correction (sc) successfully used for open-shell systems by Petersson et al. in the CBS-QB3⁵ and W1Usc⁶ procedures would compensate for the removal of the A' parameter. This gives rise to G4-sc and G4(MP2)-sc (sc for spin contamination). Finally, we explore whether the use of damping in conjunction with the A' correction can alleviate the problems associated with a spin-polarization correction that increases indefinitely. This gives rise to G4-SP (SP for spin polarization) and G4(MP2)-SP.

2. Computational Details

Standard ab initio molecular orbital theory and density-functional theory (DFT) calculations⁷ were carried out with the Gaussian 09 program.⁸ Geometries were optimized at the B3-LYP/6-31G(2df,p) level. Zero-point vibrational energies (ZPVEs) and thermal corrections to enthalpy (ΔH) at 298 K, derived from scaled (0.9854) B3-LYP/6-31G(2df,p) frequencies, were incorporated into the total energy where

appropriate. W1⁹ single-point energies were obtained using B3-LYP/6-31G(2df,p) geometries and include ZPVEs and, for the calculation of heats of formation, thermal corrections, obtained with scaled B3-LYP/6-31G(2df,p) frequencies, i.e., W1//B3-LYP/6-31G(2df,p). Unless otherwise noted, empirical parameters are optimized by minimizing the MAD for the G3/05 test set.² The HLC parameters are listed in mhartree, while relative energies are reported in kJ mol^{-1} . All reported deviations are obtained as $E(\text{calculated}) - E(\text{reference})$.

3. Results and Discussion

3.1. G4-5H and G4(MP2)-5H Procedures. The dilemma of an indefinitely increasing HLC contribution arises from the use of different parameters, A and A' , for the number of β electrons in closed- and open-shell species, respectively. The most straightforward means to correct this problem is to eliminate A' and use a single parameter A for both closed- and open-shell species. This reduces the number of empirical HLC parameters to five. Using the G3/05 test set, we have reoptimized these five parameters (namely A , B , C , D , and E) of G4 and G4(MP2), and we term the modified procedures G4-5H and G4(MP2)-5H. We have also applied these new methods to the set of 49 radical reactions of Coote (referred to as the RR49 set).⁴ The RR49 set comprises 21 radical-addition reactions and 28 hydrogen-abstraction reactions, in which W1 values are used as benchmarks. The optimized parameters and indicators of the performance of the various G4-type procedures are shown in Table 1, which also includes corresponding data for the G4-sc and G4(MP2)-sc procedures, to be discussed in Section 3.2, as well as those for G4-SP and G4(MP2)-SP to be discussed in Section 3.3.

It is apparent that the exclusion of the A' parameter from G4 leads to slightly less good agreement with the G3/05 set of experimental data, as reflected in the slightly larger value of the MAD for G4-5H compared with G4 (though the largest deviation (LD) is smaller). On the other hand, the performance for the RR49 set is improved. However, this overall improvement is not uniform across the two types of reaction. Thus, omission of A' leads to a larger MAD for addition reactions but a smaller MAD for hydrogen abstraction. This is consistent with the previous observation,⁴ that the exclusion of the HLC from G4 makes the performance worse for radical additions but leads to an improvement for hydrogen abstraction. A comparison between G4(MP2) and G4(MP2)-5H results in the same observations. Why does omission of the A' parameter lead to the deterioration for radical-addition reactions? Why does it lead to an improvement for hydrogen abstraction? To address these questions, we compare the heats of formation (ΔH_f , 298 K) calculated for the component closed-shell molecules and radicals in the RR49 set with the W1 values (Table 2).

We can see that G4 and G4-5H perform comparably for the closed-shell species. However, the exclusion of A' from G4 leads to a somewhat poorer description by G4-5H for the radicals. Both G4 and G4-5H give larger deviations from the W1 results for the reactants in the hydrogen-abstraction reactions (mainly saturated species) than for the reactants in

Table 1. HLC Parameters A, A', B, C, D, E, s, and a (mhartree) and r for Various G4-Type Procedures and Indicators of Their Performance for the Various Test Sets (kJ mol⁻¹)

	G4	G4-5H	G4-sc	G4-SP	G4(MP2)	G4(MP2)-5H	G4(MP2)-sc	G4(MP2)-SP
A	6.947	6.727	6.741	6.946	9.472	9.570	9.914	9.477
A'	7.128				9.769			
B	2.441	3.101	2.890	2.426	3.102	4.686	4.434	3.103
C	7.116	6.897	6.911	7.115	9.741	9.838	10.166	9.741
D	1.414	1.304	1.300	1.414	2.115	2.162	2.312	2.121
E	2.745	2.725	1.642	1.816	2.379	2.810	2.962	2.769
s			3.914				9.799	
a				0.202				0.329
r				0.962				0.962
G3/05 ^a								
MD ^b	-0.21	-0.17	-0.32	-0.22	-0.80	-0.69	-0.67	-0.74
MAD ^c	3.48	3.58	3.51	3.49	4.36	4.60	4.47	4.37
LD ^d	-37.30	-32.01	-31.46	-36.09	-41.61	-33.04	-31.58	-39.62
STD ^e	4.96	4.98	4.94	4.96	6.22	6.37	6.23	6.21
NO ^f	36	39	40	38	69	80	73	67
RR49 ^g								
MD	1.79	0.97	1.02	1.68	2.44	1.10	1.22	2.26
MAD	2.75	2.21	2.09	2.27	4.44	2.96	2.67	3.46
LD	5.55	8.25	8.14	4.46	8.81	8.07	7.80	6.97
STD	2.56	2.79	2.63	1.94	4.48	3.48	3.12	3.26
NO	0	0	0	0	3	0	0	0
ADD ^h								
MD	-0.39	3.69	3.51	0.34	-1.99	4.70	4.25	-0.76
MAD	1.87	3.74	3.56	1.70	2.68	4.70	4.25	2.02
LD	-4.49	8.25	8.14	4.14	-7.42	8.07	7.80	-5.58
STD	2.27	2.11	2.06	2.12	2.72	1.96	2.00	2.38
NO	0	0	0	0	0	0	0	0
ABS ⁱ								
MD	3.42	-1.07	-0.85	2.69	5.75	-1.60	-1.05	4.53
MAD	3.42	1.07	0.99	2.69	5.75	1.65	1.48	4.53
LD	5.55	-3.42	-3.32	4.46	8.81	-4.51	-4.25	6.97
STD	1.17	0.71	0.86	0.95	1.88	1.03	1.37	1.48
NO	0	0	0	0	3	0	0	0
MSM11 ^j								
MD					-11.46	11.01	8.08	-1.07
MAD					9.00	8.94	6.43	3.63
LD					-23.58	22.39	-11.97	-11.48
STD					8.23	6.46	3.93	5.72
NO					6	8	7	1

^a The G3/05 test set of ref 2. ^b Mean deviation. ^c Mean absolute deviation. ^d Largest deviation. ^e Standard deviation. ^f Number of outliers (absolute deviation > 8.4 kJ mol⁻¹ (~2 kcal mol⁻¹)). ^g The radical reaction test set of ref 4. ^h Radical-addition subset of RR49. ⁱ Hydrogen-abstraction subset of RR49. ^j Experimental IE values taken from the NIST Chemistry Webbook¹⁰ and ref 11.

the radical-addition reactions (unsaturated species). A detailed analysis of individual deviations for the radical-addition reactions (see Supporting Information, Table S2) shows that the poorer performance of G4-5H can be attributed mainly to the larger deviations for radicals. The apparent improved accuracy for hydrogen abstraction is also primarily due to the inferior description of radicals, which leads to a fortuitous cancellation of errors with those for the closed-shell reactants. We again find that the differences between G4(MP2) and G4(MP2)-5H are similar to those observed between G4 and G4-5H.

3.2. Spin-Corrected G4 and G4(MP2) Procedures.

With the benefit of knowing how the additional parameter A' affects the accuracy of G4 for different types of species, can we formulate an alternative form of parametrization, that retains the accuracy of G4 on radical species, but without the feature of a diverging HLC? One approach that has been used by Petersson et al. to account for deficiencies in the theoretical descriptions of radicals is to introduce a spin-correction (sc) term, such as the one employed in the CBS-

QB3⁵ and the W1Usc⁶ procedures. The spin-correction term (ΔE_{sc}) is given by the general formula:

$$\Delta E_{sc} = -s\Delta\langle S^2 \rangle \quad (6)$$

where s is a fitted parameter and $\Delta\langle S^2 \rangle$ is the deviation of the UHF spin-squared expectation value $\Delta\langle S^2 \rangle$ from the ideal value.

We have introduced this simple expression for the spin correction (ΔE_{sc}) in G4-5H and G4(MP2)-5H. The value of $\Delta\langle S^2 \rangle$ is evaluated with $\langle S^2 \rangle$ obtained from the UHF/aug-cc-pV5Z and UHF/aug-cc-pV(Q+d)Z calculations, for G4 and G4(MP2), respectively. The five HLC parameters A–E, as well as s , are optimized using the G3/05 set. The new procedures are termed G4-sc and G4(MP2)-sc, and have six HLC parameters.

For the G3/05 set, we see (Table 1) that the inclusion of ΔE_{sc} leads to a slightly smaller MAD than G4-5H. In a similar vein, G4(MP2)-sc also performs slightly better than G4(MP2)-5H. For the RR49 set, we again see a slightly lower

Table 2. Performance of G4-Type Procedures Compared With W1 for the Heats of Formation for Species Involved in the RR49 Set of Reactions (kJ mol⁻¹)^a

	G4	G4-5H	G4-sc	G4-SP	G4 (MP2)	G4 (MP2)-5H	G4 (MP2)-sc	G4 (MP2)-SP
Closed-Shell Reactants (X) in Radical-Addition Reactions (15)								
MD	1.54	1.53	1.58	1.55	0.55	0.50	0.52	0.59
MAD	3.01	3.01	3.01	3.01	3.61	3.59	3.54	3.62
LD	5.39	5.39	5.49	5.40	-8.82	-8.85	-8.72	-8.81
STD	3.02	3.02	3.00	3.02	4.33	4.32	4.26	4.33
NO	0	0	0	0	0	0	0	0
Closed-Shell Reactants (R2-H) in Hydrogen-Abstraction Reactions (22)								
MD ^b	4.78	4.77	4.79	4.79	5.08	5.01	4.93	5.14
MAD ^c	4.98	4.97	4.96	4.99	5.92	5.86	5.75	5.98
LD ^d	8.24	8.22	8.21	8.25	9.74	9.66	9.47	9.74
STD ^e	2.61	2.61	2.56	2.61	4.21	4.20	4.10	4.23
NO ^f	0	0	0	0	3	3	3	3
Radicals in Radical-Addition and Hydrogen-Abstraction Reactions (29)								
MD	0.30	4.02	4.30	0.95	-1.30	4.04	4.29	-0.22
MAD	1.62	4.34	4.54	1.86	2.54	4.96	5.06	2.41
LD	-7.17	7.37	7.77	-6.60	-15.77	-10.50	-9.54	-14.95
STD	2.21	2.44	2.59	2.18	3.73	3.81	3.97	3.65
NO	0	0	0	0	0	0	1	0
All Species in Radical-Addition and Hydrogen-Abstraction Reactions (66)								
MD	2.24	4.35	4.51	2.61	1.45	4.46	4.57	2.09
MAD	3.07	4.61	4.72	3.21	4.00	5.35	5.35	3.95
LD	8.24	8.22	8.21	8.25	-15.77	-10.50	-9.54	-14.95
STD	3.26	2.52	2.56	3.04	5.04	3.97	4.00	4.71
NO	0	0	0	0	3	3	4	3

^a From test set of ref 4. ^b Mean deviation. ^c Mean absolute deviation. ^d Largest deviation. ^e Standard deviation. ^f Number of outliers (absolute deviation > 8.4 kJ mol⁻¹ (~2 kcal mol⁻¹)).

MAD for G4-sc compared with G4-5H, for both the radical-addition reactions and hydrogen abstraction. There is also a more pronounced improvement for G4(MP2)-sc over G4(MP2)-5H. Thus, inclusion of the ΔE_{sc} term leads to a consistent improvement for both G4-5H and G4(MP2)-5H.

Despite this improvement, we find that G4-sc and G4(MP2)-sc still perform slightly less well than standard G4 and G4(MP2) for the G3/05 test set. On the other hand, the sc variants show significantly improved overall performance for the RR49 set compared with G4 and G4(MP2). When we inspect the component species in the RR49 reactions (Table 2), we see that, like G4-5H and G4(MP2)-5H, the improvement in the reaction energies is partly the result of a fortuitous cancellation of errors, rather than a broad improvement for all species. Thus, while G4-sc performs comparably to G4 for closed-shell molecules, it deviates more substantially from W1 than does G4 for radicals. Similar features are observed for G4(MP2)-sc versus G4(MP2).

3.3. Formulation of the G4-SP and G4(MP2)-SP Procedures. We have seen in the last section that the “sc” procedures provide an alternative to G4 and G4(MP2), with an accuracy that approaches that of the original methods but without the shortcoming of a divergent A' correction term. In this section, we derive a second type of alternative, represented by G4-SP and G4(MP2)-SP. We aim to further close the gap in performance with G4 and G4(MP2), in addition to eliminating the drawback of a diverging HLC.

In the original G4 procedure,^{1d} the A' parameter was introduced to account for the effect of spin polarization by the radical center to the rest of the molecule. As such, one may consider A' as a perturbed version of the parameter A :

$$A' = A + x \quad (7)$$

Thus, the HLC for radicals can be written as:

$$-(A + x)n_{\beta} - B(n_{\alpha} - n_{\beta}) \quad (8)$$

The problem that exists with G4 is that the correction term increases linearly with n_{β} , i.e., each of the β electrons has the same individual effect in a large radical as they do in a small radical, and so, the total effect for n_{β} electrons, i.e., $-xn_{\beta}$ or $-(A' - A)n_{\beta}$, is likely to be overestimated. A more reasonable physical picture comes about by allowing the effect of successive electrons to be damped. This can be accomplished using a geometric series with ratio $r < 1$ and leads to contributions to xn_{β} of the individual n_{β} electrons in the HLC (eq 8) being of the form:

$$ar^{m-1} \quad (m = 1, \dots, n_{\beta}) \quad (9)$$

Standard G4 corresponds to $a = A' - A$ and $r = 1$, with each of the n_{β} electrons making an equal contribution of $A - A'$.

The sum for n_{β} electrons (for $r \neq 1$) is then:

$$xn_{\beta} = \sum_{m=1}^{n_{\beta}} ar^{m-1} = \frac{a(1 - r^{n_{\beta}})}{1 - r} \quad (10)$$

In this formulation, if $r < 1$, then as n_{β} becomes large the individual contributions of the β electrons (eq 9) approach zero and the sum (eq 10) approaches the limiting value:

$$\frac{a}{1 - r} \quad (11)$$

For radicals with no β electrons, i.e., $n_{\beta} = 0$, there is no correction for spin polarization, as should be the case. The

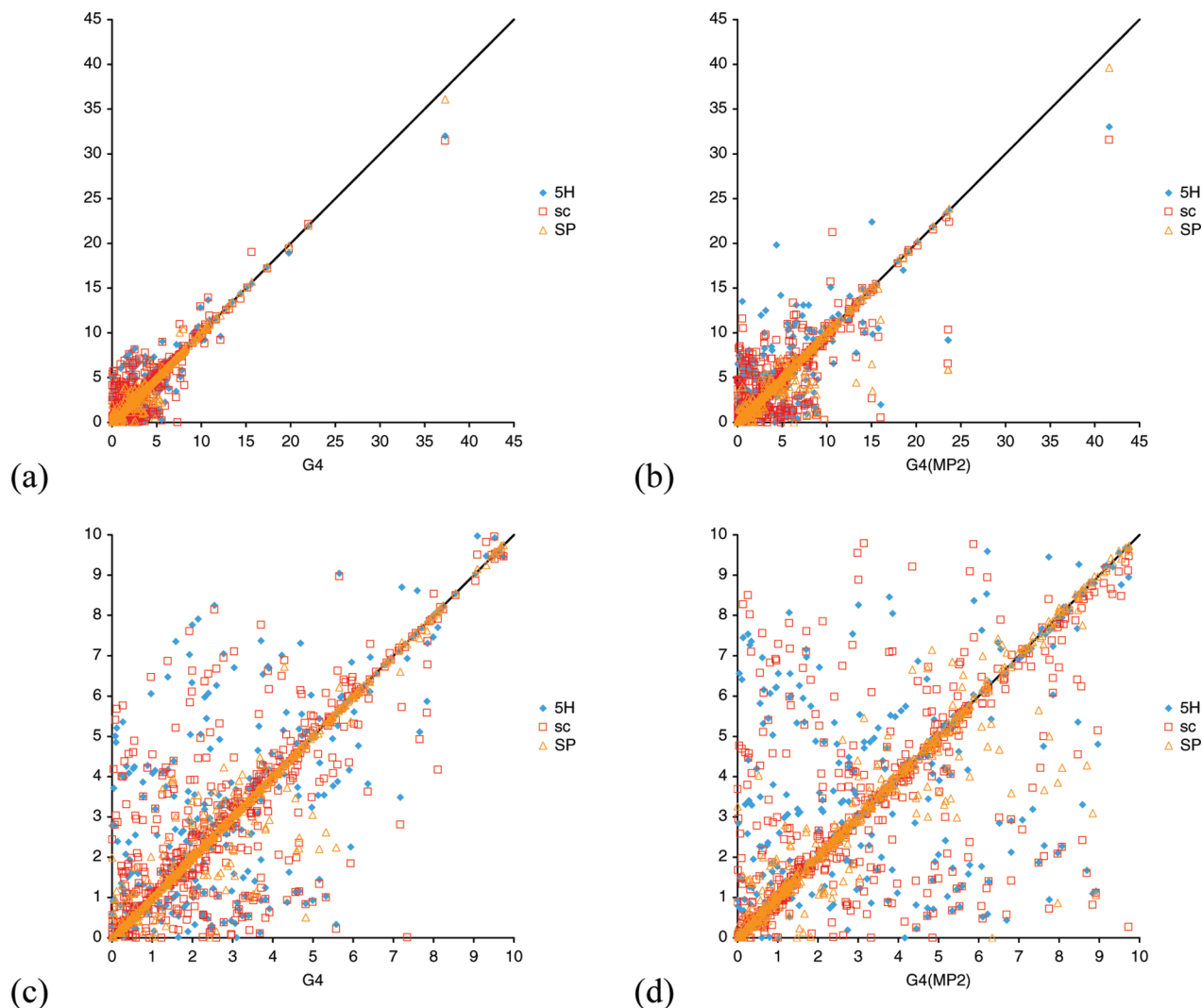


Figure 1. Absolute deviations (kJ mol^{-1}) for the modified versus the original procedures for G4 (G3/05 and RR49 sets) and G4(MP2) (G3/05, RR49, and MSM11 sets). The full set of results are presented in (a) (G4) and (b) (G4(MP2)), while expanded displays of the region up to 10 kJ mol^{-1} are presented in (c) (G4) and (d) (G4(MP2)).

term a represents the correction for the most proximate electron, while the parameter r governs the rate of decay for such a correction. Together, they determine the maximum correction for spin polarization in a radical. The total number of HLC parameters for G4-SP and G4(MP2)-SP is then seven.

In order to obtain appropriate optimum values for a and r , we need to include thermochemical data for larger molecules to allow for a representation of the converging behavior of spin polarization. To this end, we have added ionization energy (IE) data for a total of 11 larger molecules to the G3/05 set. These species include four polycyclic aromatic hydrocarbons (PAH): naphthalene (C_{10}H_8), anthracene ($\text{C}_{14}\text{H}_{10}$), pyrene ($\text{C}_{16}\text{H}_{10}$), and coronene ($\text{C}_{24}\text{H}_{12}$); three diamondoids: adamantane ($\text{C}_{10}\text{H}_{16}$), diadamantane ($\text{C}_{14}\text{H}_{20}$), and triadamantane ($\text{C}_{18}\text{H}_{24}$); two perfluorocarbons (PFC): perfluorobenzene (C_6F_6) and perfluoronaphthalene (C_{10}F_8); and finally CCl_4 and SF_6 . We term this collection of IEs the medium-sized-molecule (MSM11) set and the enlarged G3/05 set G3/05+. Optimization for G4(MP2) over this larger set yields $r = 0.962$.

Whereas all the parameters for G4(MP2)-SP are optimized using the full G3/05+ set, single-point G4 calculations for several of the MSM11 molecules are not computationally tractable for us at the present time. Consequently, to obtain the parameters for G4-SP, we assume that the value for r in G4-SP is the same as that in G4(MP2)-SP, i.e., 0.962. The HLC parameters A–E and a are then optimized using G3/05.

Having obtained the optimized parameters, we can describe the protocol for G4-SP and G4(MP2)-SP as follows: (1) Geometry optimization and harmonic frequency calculations are performed at the B3-LYP/6-31G(2df,p) level. ZPVEs and thermal corrections are obtained using frequencies scaled by 0.9854. (2) A series of single-point energies are obtained to approximate the CCSD(T)/CBS energy using composite schemes detailed in the G4^{1d} and G4(MP2)^{3c} papers. (3) An empirical higher-level correction (HLC) is added to the total energy. This takes the same form as detailed in G4 and G4(MP2), with the exception of open-shell molecular species, for which the (n_α, n_β) component of the HLC term is obtained as $-An_\beta - B(n_\alpha - n_\beta)$. The new

values for these parameters are included in Table 1. (4) In addition, for open-shell molecular species, a spin-polarization (SP) correction term is added to the total energy. This takes the form $a(1 - r^{n_{\beta}})/(1 - r)$, where n_{β} is the number of β electrons, $a = 0.202$ and 0.329 mhartrees for G4-SP and G4(MP2)-SP, respectively, and $r = 0.962$ for both procedures.

3.4. Performance of G4-SP and G4(MP2)-SP. How do the “SP” procedures compare with the “sc” procedures? We can see from Table 1 that the SP procedures give slightly lower MAD values for the G3/05 set, with statistical indicators very similar to the original G4 and G4(MP2). For the RR49 set and its subsets, G4-SP appears to be superior to G4, and G4(MP2)-SP also compares favorably to G4(MP2). G4-sc and G4(MP2)-sc perform even better for the overall RR49 set, due to a smaller MAD for the abstraction reactions. However, it is evident from Table 2 that this is mainly due to the poorer description of the radicals, rather than to an improved performance for the closed-shell species. Comparison of the various G4(MP2) procedures for the IEs of the larger molecules (Table 1) shows quite striking results. We find that G4(MP2)-sc performs somewhat better than G4(MP2) and G4(MP2)-5H, while a more pronounced improvement can be seen for the G4(MP2)-SP procedure.

Figure 1 displays plots of the absolute deviations (from experiment or W1) of the various modified procedures against results for standard G4 and G4(MP2). It is apparent that, for most of the data, the “SP” procedures perform very similarly to G4 and G4(MP2), while they show improvement in a number of cases. Overall, the “5H” and “sc” procedures do not perform quite as well. There are somewhat more cases where these two types of modification lead to larger deviations than G4 and G4(MP2), i.e., they lie above the line, than cases where smaller deviations are achieved. Thus, we recommend the general use of G4-SP and G4(MP2)-SP as alternatives to G4 and G4(MP2). Nonetheless, the use of the “sc” methods may be beneficial for specific applications in which cancellation of errors may lead to closer agreement with experiment.

4. Concluding Remarks

In the standard G4 and G4(MP2) procedures, the difference in the treatment of closed-shell and open-shell species, by the use of different values for the HLC parameters for A (closed-shell) and A' (open-shell), can lead to shortcomings in the description of radical reactions. This pitfall is most evident when large radicals are involved and arises because of an indefinitely increasing HLC contribution to the reaction energy with increasing size of radical.

We have formulated a number of minor modifications to these procedures to overcome such a deficiency. One approach is simply to remove the A' parameter, i.e., make $A' = A$. These methods are termed G4-5H and G4(MP2)-5H and include five HLC parameters. Overall, they perform slightly worse than G4 and G4(MP2) but without the A' problem for large radicals.

A second approach is to remove the A' parameter but to include a spin-correction term instead. These methods are termed G4-sc and G4(MP2)-sc and include six HLC parameters. They again perform only slightly worse than G4 and

G4(MP2) overall but without the A' problem for large radicals. They are slightly better than G4-5H and G4(MP2)-5H. The use of G4-sc and G4(MP2)-sc may be beneficial for specific applications in which cancellation of errors may lead to closer agreement with the experiment.

A third type of alternative method, represented by G4-SP and G4(MP2)-SP, also equates A and A'. In this case, the spin-polarization effect, that is originally treated by having different A and A' values, is handled by a new damped SP term. This term is derived from a geometric series and converges to a predefined maximum. Hence, it does not have the dilemma of significant overestimation of spin polarization for large radicals. We have demonstrated that the “SP” formulation, which includes seven HLC parameters, performs somewhat better than the “sc” procedures. The SP procedures work as well as G4 and G4(MP2) for small molecules, while appearing to be somewhat more accurate for medium-sized systems.

Acknowledgment. We gratefully acknowledge valuable assistance from Dr. Paul. C. Redfern and Professor Larry A. Curtiss, the award of an Australian Professorial Fellowship (to L.R.), funding from the ARC Centre of Excellence for Free Radical Chemistry and Biotechnology, and generous allocations of computer time from the National Computational Infrastructure (NCI) National Facility, Intersect, and the Australian Centre for Advanced Computing and Communications (ac3).

Supporting Information Available: Zero-point vibrational energies and thermal corrections from scaled B3-LYP/6-31G(2df,p) frequencies and G4-5H, G4-sc, G4-SP, G4(MP2)-5H, G4(MP2)-sc, and G4(MP2)-SP total electronic energies (Table S1) and deviations from experimental and W1 energies (Table S2). This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- (1) (a) Pople, J. A.; Head-Grodon, M.; Fox, D. J.; Raghavachari, K.; Curtiss, L. A. *J. Chem. Phys.* **1989**, *90*, 5622. (b) Curtiss, L. A.; Raghavachari, K.; Trucks, G. W.; Pople, J. A. *J. Chem. Phys.* **1991**, *94*, 7221. (c) Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Rassolov, V.; Pople, J. A. *J. Chem. Phys.* **1998**, *109*, 7764. (d) Curtiss, L. A.; Redfern, P. C.; Raghavachari, K. *J. Chem. Phys.* **2007**, *126*, 084108.
- (2) Curtiss, L. A.; Redfern, P. C.; Raghavachari, K. *J. Chem. Phys.* **2005**, *123*, 124107.
- (3) (a) Curtiss, L. A.; Raghavachari, K.; Pople, J. A. *J. Chem. Phys.* **1993**, *98*, 1293. (b) Curtiss, L. A.; Redfern, P. C.; Raghavachari, K.; Rassolov, V.; Pople, J. A. *J. Chem. Phys.* **1999**, *110*, 4703. (c) Curtiss, L. A.; Redfern, P. C.; Raghavachari, K. *J. Chem. Phys.* **2007**, *127*, 124105.
- (4) Lin, C. Y.; Hodgson, J. L.; Namazian, M.; Coote, M. L. *J. Phys. Chem. A* **2009**, *113*, 3690.
- (5) (a) Montgomery, J. A., Jr.; Frisch, M. J.; Ochterski, J. W.; Petersson, G. A. *J. Chem. Phys.* **1999**, *110*, 2822. (b) Montgomery, J. A., Jr.; Frisch, M. J.; Ochterski, J. W.; Petersson, G. A. *J. Chem. Phys.* **2000**, *112*, 6532.
- (6) (a) Barnes, E. C.; Petersson, G. A.; Montgomery, J. A.; Frisch, M. J.; Martin, J. M. L. *J. Chem. Theory Comput.* **2009**, *5*, 2687. (b) Wood, G. P. F.; Radom, L.; Petersson, G. A.; Barnes,

- E. C.; Frisch, M. J.; Montgomery, J. A., Jr. *J. Chem. Phys.* **2006**, *125*, 094106.
- (7) See, for example: (a) Hehre, W. J.; Radom, L.; Schleyer, P. v. P.; Pople, J. A. *Ab Initio Molecular Orbital Theory*; Wiley: New York, 1986. (b) Koch, W.; Holthausen, M. C. *A Chemist's Guide to Density Functional Theory*, 2nd ed.; Wiley: New York, 2001. (c) Jensen, F. *Introduction to Computational Chemistry*, 2nd ed.; Wiley: Chichester, 2007.
- (8) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R. E.; Stratmann, O.; Yazyev, A. J.; Austin, R.; Cammi, C.; Pomelli, J. W.; Ochterski, R.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, O.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. *Gaussian 09*, Revision A.02, Gaussian, Inc.: Wallingford CT, 2009.
- (9) Martin, J. M. L.; de Oliveira, G. *J. Chem. Phys.* **1999**, *111*, 1843.
- (10) Linstrom, P. J., Mallard, W. G., Eds. *NIST Chemistry WebBook*; NIST Standard Reference Database Number 69; National Institute of Standards and Technology: Gaithersburg MD, 20899, <http://webbook.nist.gov> (accessed May 2010).
- (11) Lenzke, K.; Landt, L.; Hoener, M.; Thomas, H.; Dahl, J. E.; Liu, S. G.; Carlson, R. M. K.; Möller, T.; Bostedt, C. *J. Chem. Phys.* **2007**, *127*, 084320.

CT100266U

JCTC

Journal of Chemical Theory and Computation

A Parallel Iterative Method for Computing Molecular Absorption Spectra

Peter Koval,^{*,†} Dietrich Foerster,[‡] and Olivier Coulaud[§]

CNRS, HiePACS project, LaBRI, 351 Cours de la Liberation, 33405, Talence, France,
CPMOH, University of Bordeaux 1, 351 Cours de la Liberation, 33405, Talence, France,
and INRIA SUD OUEST, HiePACS project, 351 Cours de la Liberation,
33405, Talence, France

Received May 28, 2010

Abstract: We describe a fast parallel iterative method for computing molecular absorption spectra within TDDFT linear response and using the LCAO method. We use a local basis of “dominant products” to parametrize the space of orbital products that occur in the LCAO approach. In this basis, the dynamic polarizability is computed iteratively within an appropriate Krylov subspace. The iterative procedure uses a matrix-free GMRES method to determine the (interacting) density response. The resulting code is about 1 order of magnitude faster than our previous full-matrix method. This acceleration makes the speed of our TDDFT code comparable with codes based on Casida’s equation. The implementation of our method uses hybrid MPI and OpenMP parallelization in which load balancing and memory access are optimized. To validate our approach and to establish benchmarks, we compute spectra of large molecules on various types of parallel machines. The methods developed here are fairly general, and we believe they will find useful applications in molecular physics/chemistry, even for problems that are beyond TDDFT, such as organic semiconductors, particularly in photovoltaics.

1. Introduction

The standard way to investigate the electronic structure of matter is by measuring its response to external electromagnetic fields. To describe the electronic response of molecules, one may use time-dependent density functional theory (TDDFT).¹ TDDFT has been particularly successful in the calculation of absorption spectra of finite systems such as atoms, molecules, and clusters,^{1,2} and for such systems it remains the computationally cheapest *ab initio* approach without any empirical parameters.

In the framework of TDDFT,³ time-dependent Kohn–Sham-like equations replace the Kohn–Sham equations of the static density-functional theory (DFT). Although these equations can be applied to very general situations,⁴ we will restrict ourselves to the case where the external interaction

with light is small. This condition is satisfied in most practical applications of spectroscopy and can be treated by the linear response approximation.

TDDFT focuses on the time-dependent electron density $n(\mathbf{r}, t)$. One assumes the existence of an artificial time-dependent potential $V_{KS}(\mathbf{r}, t)$ in which (equally artificial) noninteracting reference electrons acquire exactly the same time-dependent density $n(\mathbf{r}, t)$ as the interacting electrons under study. The artificial time-dependent potential $V_{KS}(\mathbf{r}, t)$ is related to the true external potential $V_{ext}(\mathbf{r}, t)$ by the following equation:

$$V_{KS}(\mathbf{r}, t) = V_{ext}(\mathbf{r}, t) + V_H(\mathbf{r}, t) + V_{xc}(\mathbf{r}, t)$$

Here, $V_H(\mathbf{r}, t)$ is the Coulomb potential of the electronic density $n(\mathbf{r}, t)$, and $V_{xc}(\mathbf{r}, t)$ is the exchange-correlation potential. The exchange-correlation potential absorbs all the nontrivial dynamics, and in practice it is usually taken from numerical studies of the interacting homogeneous electron gas.

* To whom correspondence should be addressed E-mail: koval.peter@gmail.com.

† CNRS, HiePACS project, LaBRI.

‡ CPMOH, University of Bordeaux 1.

§ INRIA SUD OUEST, HiePACS project.

The above time-dependent extension of the Kohn–Sham equation was used by Gross et al.⁵ to find the dynamic linear response function $\chi = \delta n(\mathbf{r}, t)/\delta V_{\text{ext}}(\mathbf{r}', t')$ of an interacting electron gas, a response function that expresses the change of the electron density $\delta n(\mathbf{r}, t)$ upon an external perturbation $\delta V_{\text{ext}}(\mathbf{r}', t')$. From the change $\delta n(\mathbf{r}, t)$ of the electron density, one may calculate the polarization $\delta \mathbf{P} = \int \delta n(\mathbf{r}, t) \mathbf{r} d^3r$ that is induced by an external field $\delta V_{\text{ext}}(\mathbf{r}, t)$. The imaginary part of its Fourier transform $\delta \mathbf{P}(\omega)$ provides us with the absorption coefficient, and the poles of its Fourier transform provide information on electronic transitions.

Most practical implementations of TDDFT in the regime of linear response are based on Casida's equations.^{6,7} Casida has derived a Hamiltonian in the space of particle-hole excitations, the eigenstates of which correspond to the poles of the interacting response function $\chi(\mathbf{r}, \mathbf{r}', \omega)$. Although Casida's approach has an enormous impact in chemistry,⁷ it is computationally demanding for large molecules. This is so because Casida's Hamiltonian acts in the space of particle-hole excitations, the number of which increases as the square of the number of atoms.

Alternatively, one may also solve the TDDFT linear response by iterative methods. A first example of an iterative method for computing molecular spectra is the direct calculation of the density change $\delta n(\mathbf{r}, \omega)$ in a modified Sternheimer approach.^{1,7} In this scheme, one determines the variation of the Kohn–Sham orbitals $\delta \psi(\mathbf{r}, t)$ due to an external potential without using any response functions. By contrast, use of the response functions allowed van Gisbergen et al.^{8,9} to develop a self-consistent iterative procedure for the variation of the density $\delta n(\mathbf{r}, t)$ without invoking the variation of the molecular orbitals. The absorption and dichroism properties have been extensively calculated with the response function¹⁰ including iterative calculations. There exists also an iterative approach based on the density matrix.^{11,12} This approach is quite different from ours, but its excellent results, obtained with a plane-wave basis, serve as a useful test of our LCAO-based method.

Over the past few years, the authors of the present paper developed and applied a new basis in the space of products that appear in the application of the LCAO method to excited states.^{13–15} This methodological improvement allows for a simplified iterative approach for computing molecular spectra, and the present paper describes this approach. We believe that the methods developed here will be useful in molecular physics, not only in the context of TDDFT but also for systems such as large organic molecules that, because of their excitonic features, require methods beyond ordinary TDDFT.

This paper is organized as follows: In section 2, we briefly recall the TDDFT linear response equations. In section 3, we introduce a basis in the space of products of atomic orbitals. In section 4, we describe our iterative method of computing the polarizability, and in section 5, we explain how the interaction kernels are computed. Section 6 describes the parallelization of our iterative method in both multithread and multiprocessing modes. In section 7, we present results and benchmarks of the parallel implementation of our code. We conclude in section 8.

2. Brief Review of Linear Response in TDDFT

To find the equations of the TDDFT linear response, one starts from the time-dependent extension of the Kohn–Sham equation that was already mentioned in the introduction and which we rewrite as follows:

$$V_{\text{KS}}([n], \mathbf{r}, t) = V_{\text{ext}}([n], \mathbf{r}, t) + V_{\text{H}}([n], \mathbf{r}, t) + V_{\text{xc}}([n], \mathbf{r}, t)$$

The notation $[n]$ indicates that the potentials V_{KS} , V_{H} , and V_{xc} in this equation depend on the distribution of electronic charge $n(\mathbf{r}, t)$ in all of space and at all times. To find out how the terms of this equation respond to a small variation, $\delta n(\mathbf{r}, t)$, of the electron density, we take their variational derivative with respect to the electron density:

$$\frac{\delta V_{\text{KS}}([n], \mathbf{r}, t)}{\delta n(\mathbf{r}', t')} = \frac{\delta V_{\text{ext}}([n], \mathbf{r}, t)}{\delta n(\mathbf{r}', t')} + \frac{\delta}{\delta n(\mathbf{r}', t')} [V_{\text{H}}([n], \mathbf{r}, t) + V_{\text{xc}}([n], \mathbf{r}, t)]$$

The important point to notice here is that $\delta V_{\text{KS}}/\delta n$ and $\delta V_{\text{ext}}/\delta n$ are inverses of $\chi_0 = \delta n/\delta V_{\text{KS}}$ and $\chi = \delta n/\delta V_{\text{ext}}$ that represent the response of the density of free and interacting electrons to, respectively, variations of the potentials V_{KS} and V_{ext} . As the response χ_0 of free electrons is known and since we wish to find the response χ of interacting electrons, we rewrite the previous equation compactly as follows:

$$\chi^{-1} = \chi_0^{-1} - f_{\text{Hxc}} \quad (1)$$

Here, an interaction kernel $f_{\text{Hxc}} = \delta/\delta n[V_{\text{H}} + V_{\text{xc}}]$ is introduced. Equation 1 has the form of a Dyson equation that is familiar from many-body perturbation theory.¹⁶

To make use of eq 1, it remains to specify χ_0 and f_{Hxc} . The Kohn–Sham response function χ_0 can be computed in terms of orbitals and eigenenergies,¹ as will be discussed later in this section (see eq 3 below). The exchange–correlation kernel $\delta V_{\text{xc}}/\delta n$ has to be approximated in practice. In this work, we use the simplest adiabatic local density approximation (ALDA) for the exchange–correlation potential $V_{\text{xc}}([n], \mathbf{r}, t) = V_{\text{xc}}(n(\mathbf{r}, t))$, and the kernel f_{Hxc} becomes:

$$f_{\text{Hxc}} = \frac{\delta(t - t')}{|\mathbf{r} - \mathbf{r}'|} + \delta(\mathbf{r} - \mathbf{r}') \delta(t - t') \frac{dV_{\text{xc}}}{dn}$$

Moreover, the exact-exchange functional can be easily formulated on the basis of dominant products, although this is beyond the scope of the present work.

In principle one could determine the interacting response function $\chi(\mathbf{r}, t, \mathbf{r}', t')$ from eq 1, determine the variation of density $\delta n(\mathbf{r}, t) = \int \chi(\mathbf{r}, t, \mathbf{r}', t') \delta V_{\text{ext}}(\mathbf{r}', t') d^3r' dt'$ due to a variation of the external potential δV_{ext} and find the observable polarization $\delta \mathbf{P} = \int \delta n(\mathbf{r}, t) \mathbf{r} d^3r$. To realize these operations in practice, we use a basis of localized functions to represent the space dependence of the functions $\chi_0(\mathbf{r}, \mathbf{r}', \omega)$ and $f_{\text{Hxc}}(\mathbf{r}, \mathbf{r}')$.

In order to introduce such a basis into the Petersilka–Gossmann–Gross equation, eq 1, we eliminate the inversions in this equation and transform to the frequency domain:

$$\chi(\mathbf{r}, \mathbf{r}', \omega) = \chi_0(\mathbf{r}, \mathbf{r}', \omega) + \int d^3 r_1 d^3 r_2 \chi_0(\mathbf{r}, \mathbf{r}_1, \omega) f_{\text{Hxc}}(\mathbf{r}_1, \mathbf{r}_2) \chi(\mathbf{r}_2, \mathbf{r}', \omega) \quad (2)$$

The density response function of free electrons $\chi_0(\mathbf{r}, \mathbf{r}', \omega)$ can be expressed¹ in terms of molecular Kohn–Sham orbitals as follows:

$$\chi_0(\mathbf{r}, \mathbf{r}', \omega) = \sum_{E,F} (n_F - n_E) \frac{\psi_E(\mathbf{r}) \psi_F(\mathbf{r}) \psi_F(\mathbf{r}') \psi_E(\mathbf{r}')}{\omega - (E - F) + i\varepsilon} \quad (3)$$

Here, n_E and $\psi_E(\mathbf{r})$ are occupation factors and Kohn–Sham eigenstates of eigenenergy E , and the constant ε regularizes the expression, giving rise to a Lorentzian shape of the resonances. The eigenenergies E and F are shifted in such a way that occupied and virtual states have, respectively, negative and positive energies. Only pairs E, F of opposite signs, $E \cdot F < 0$, contribute in the summation, as is appropriate for transitions from occupied to empty states.

We express molecular orbitals $\psi_E(\mathbf{r})$ as linear combinations of atomic orbitals (LCAO):

$$\psi_E(\mathbf{r}) = X_a^E f^a(\mathbf{r}) \quad (4)$$

where $f^a(\mathbf{r})$ is an atomic orbital. The coefficients X_a^E are determined by diagonalizing the Kohn–Sham Hamiltonian that is the output of a prior DFT calculation, and we assume these quantities to be available. For the convenience of the reader, we use Einstein's convention of summing over repeated indices.

3. Treatment of Excited States within LCAO

The LCAO method was developed in the early days of quantum mechanics to express molecular orbitals as linear combinations of atomic orbitals. When inserting the LCAO ansatz (eq 4) into eq 3 to describe the density response, one encounters products of localized functions $f^a(\mathbf{r}) f^b(\mathbf{r})$ —a set of quantities that are known to be linearly dependent.

There is extensive literature^{17–20} on the linear dependence of products of atomic orbitals. Baerends et al. use an auxiliary basis of localized functions to represent the electronic density.^{20,21} Their procedure of fitting densities by auxiliary functions is essential both for solving Casida's equations and in van Gisbergen's iterative approach.

In the alternative approach of Beebe and Linderberg,¹⁷ one forms the overlaps of products $\langle ab|a'b' \rangle$ to disentangle the linear dependence of the products $f^a(r) f^b(r)$. The difficulty with this approach is its lack of locality and the $O(N^4)$ cost of the construction of the overlaps.²²

Our approach is applicable to numerically given atomic orbitals of finite support, and in the special case of products on coincident atoms, it resembles an earlier construction by Aryasetiawan and Gunnarsson in the muffin tin context.²³ We ensure locality by focusing on the products of atomic orbitals for each overlapping pair of atoms at a time.^{13,14} Because our construction is local in space, it requires only $O(N)$ operations. Our procedure removes a substantial part of the linear dependence from the set of products $\{f^a(\mathbf{r}) f^b(\mathbf{r})\}$. As a result, we find a vertex-like identity for the original products in terms of certain “dominant products” $F^\lambda(\mathbf{r})$:

$$f^a(\mathbf{r}) f^b(\mathbf{r}) \sim \sum_{\lambda > \lambda_{\min}} V_\lambda^{ab} F^\lambda(\mathbf{r}) \quad (5)$$

Here, the notation V_λ^{ab} alludes to the fact that the vertex, V_λ^{ab} , was obtained from an eigenvalue problem for the pair of atoms that the orbitals a and b belong to. The condition $\lambda > \lambda_{\min}$ says that the functions corresponding to the (very many) small eigenvalues were discarded. By construction, the vertex V_λ^{ab} is nonzero only for a few orbitals, a and b , that refer to the same pair of atoms that λ belongs to, and therefore, V_λ^{ab} is a sparse object. Empirically, the error in the representation (eq 5) vanishes exponentially fast²⁴ with the number of eigenvalues that are retained. The convergence is fully controlled by choosing the threshold for eigenvalues λ_{\min} , and from now on we assume equality in relation 5.

We introduce matrix representations of the response functions χ_0 and χ in the basis of the dominant products $\{F^\mu(\mathbf{r})\}$ as follows:

$$\begin{aligned} \chi_0(\mathbf{r}, \mathbf{r}', \omega) &= \sum_{\mu\nu} F^\mu(\mathbf{r}) \chi_{\mu\nu}^0(\omega) F^\nu(\mathbf{r}') \\ \chi(\mathbf{r}, \mathbf{r}', \omega) &= \sum_{\mu\nu} F^\mu(\mathbf{r}) \chi_{\mu\nu}(\omega) F^\nu(\mathbf{r}') \end{aligned} \quad (6)$$

For the noninteracting response, $\chi_{\mu\nu}^0(\omega)$, one has an explicit expression

$$\chi_{\mu\nu}^0(\omega) = \sum_{abcd, E, F} (n_F - n_E) \frac{(X_a^E V_\mu^{ab} X_b^F)(X_c^F V_\nu^{cd} X_d^E)}{\omega - (E - F) + i\varepsilon} \quad (7)$$

which can be obtained by inserting the LCAO ansatz (eq 4) and the vertex ansatz (eq 5) into eq 3.

We insert the expansions (eq 6) into the Dyson equation, eq 2, and obtain the Petersilka–Gossmann–Gross equation in matrix form:

$$\chi_{\mu\nu}(\omega) = \chi_{\mu\nu}^0(\omega) + \chi_{\mu\mu'}^0(\omega) f_{\text{Hxc}}^{\mu'\nu'}(\omega) \chi_{\nu'\nu}(\omega) \quad (8)$$

with the kernel $f_{\text{Hxc}}^{\mu\nu}$ defined as

$$f_{\text{Hxc}}^{\mu\nu} = \int d^3 r d^3 r' F^\mu(\mathbf{r}) f_{\text{Hxc}}(\mathbf{r}, \mathbf{r}') F^\nu(\mathbf{r}') \quad (9)$$

The calculation of this matrix will be discussed in section 5.

Since molecules are small compared to the wavelength of light, one may use the dipole approximation and define the polarizability tensor

$$P_{ik}(\omega) = \int d^3 r d^3 r' \mathbf{r}_i \mathbf{r}'_k \chi(\mathbf{r}, \mathbf{r}', \omega)$$

Moreover, using eq 8, we find an explicit expression for the interacting polarizability tensor $P_{ik}(\omega)$ in terms of the known matrices χ^0 and f_{Hxc} :

$$P_{ik}(\omega) = d_i^\mu \left(\frac{1}{1 - \chi^0(\omega) f_{\text{Hxc}}}_{\mu\nu} \right) \chi_{\nu\nu'}^0(\omega) d_k^{\nu'} \quad (10)$$

where the dipole moment has been introduced $d_i^\mu = \int f F^\mu(\mathbf{r}) \mathbf{r}_i d^3 r$. Our iterative procedure for computing molecular absorption spectra is based on eq 10. Below, we will

omit Cartesian indices i and k because we compute tensor components $P_{ik}(\omega)$ independently of each other.

4. An Iterative Method for the Calculation of the Dynamical Polarizability

In eq 10, the polarizability $P(\omega)$ is expressed as a certain average of the inverse of the matrix $A_\mu^v \equiv \delta_\mu^v - \chi_{\mu\nu}^0(\omega) f_{\text{Hxc}}^{v\nu}$. A direct inversion of the matrix A_μ^v is straightforward at this point, but it is computationally expensive and of unnecessary (machine) precision. Fortunately, iterative methods are available that provide the desired result at much lower computational cost and with a sufficient precision.

In fact, we already used an iterative biorthogonal Lanczos method to create an approximate representation of the matrix A_μ^v , which can be easily inverted within the Krylov subspace under consideration.^{14,15} In this work, however, we use a simpler approach of better efficiency to compute the polarizability (eq 10). First, we calculate an auxiliary vector $\mathbf{X} = \mathbf{A}^{-1} \chi^0 d$ by solving a linear system of equations $\mathbf{A}\mathbf{X} = \chi^0 d$ with an iterative method of the Krylov type. Second, we compute the dynamic polarizability as a scalar product $P = d^\mu X_\mu$. In this way, we avoid the construction of the full and computationally expensive response matrix χ^0 . Below, we give details on this iterative procedure.

4.1. The GMRES Method for the Iterative Solution of a Linear System of Equations. As we explained above, the polarizability $P(\omega)$ can be computed separately for each frequency, by solving the system of linear equations:

$$(1 - \chi^0(\omega) f_{\text{Hxc}}) \mathbf{X}(\omega) = \chi^0(\omega) d \quad (11)$$

and by computing the dynamic polarizability as a scalar product $P(\omega) = d^\mu X_\mu(\omega)$.

We apply a generalized minimal residual method (GMRES)^{25,26} to solve the linear system of eq 11, which is of the form $\mathbf{A}\mathbf{X} = b$. GMRES belongs to the Krylov-type methods^{25,27} that represent a large matrix \mathbf{A} in an iteratively built up Krylov-type basis $|0\rangle, |1\rangle, \dots, |i\rangle$. The first vector $|0\rangle$ in the basis is chosen equal to $|b\rangle$, while further vectors are computed recursively via $|i\rangle = \mathbf{A}|i-1\rangle$. As the vectors $|i\rangle = \mathbf{A}^i|0\rangle$ are not mutually orthogonal, one may enforce their orthogonality by using the Gram-Schmidt method:

$$|i\rangle = \mathbf{A}|i-1\rangle - \sum_{j=0}^{i-1} |j\rangle \langle j|\mathbf{A}|i-1\rangle \quad (12)$$

The orthonormal basis built in this way is used in the GMRES method to approximately solve the linear system of equations $\mathbf{A}\mathbf{X} = |b\rangle$ by minimizing the residual $|r\rangle = \mathbf{A}\mathbf{X} - |b\rangle$ within the Krylov-type subspace (eq 12). The minimization of the residual occurs when the equation $\sum_j \langle i|\mathbf{A}|j\rangle \langle j|x\rangle = \langle i|b\rangle$ is satisfied and this set of equations is of much smaller size than the original problem. When the solution in the Krylov subspace $\langle i|x\rangle$ is found, then an approximate solution in the original space can be computed from $|\mathbf{X}\rangle = \sum_i |i\rangle \langle i|x\rangle$.

A suitable stopping criterion is essential for our method, and we tested several criteria in order to keep the number of iterations small and achieve a reliable result at the same

time. The conventionally used criterion that $\varepsilon_r = |r|/|b|$ should be small is unreliable when the tolerance threshold is comparatively large ($\varepsilon_r \approx 1\%$). Therefore, we suggest an alternative combined criterion.

A natural stopping criterion for an iterative solver of the linear system of equations $\mathbf{A}\mathbf{X} = |b\rangle$ is a condition on the relative error of the solution $\varepsilon_X = |\Delta\mathbf{X}|/|\mathbf{X}|$.

In our case, the quantity of interest is the dynamic polarizability $P = \langle d|\mathbf{X}\rangle$; therefore it is meaningful to impose a limit on the relative error of the polarizability $\varepsilon_P = |\Delta P|/|P|$. Estimations of the errors ε_X and ε_P can be easily obtained because $|\Delta\mathbf{X}\rangle = \mathbf{A}^{-1}|r\rangle$ and $\Delta P = \langle d|\mathbf{A}^{-1}|r\rangle$. We estimate $|\Delta\mathbf{X}\rangle$ and ΔP using a matrix norm,²⁸ $|\Delta\mathbf{X}\rangle \approx \|\mathbf{A}^{-1}\| |r\rangle$ and $\Delta P \approx \|\mathbf{A}^{-1}\| \langle d|r\rangle$. We used the Frobenius norm of the Krylov representation of the matrix \mathbf{A}^{-1} , $\|\mathbf{A}^{-1}\| \approx (\sum_{ij} |\langle i|\mathbf{A}^{-1}|j\rangle|^2)^{1/2}$ because \mathbf{A} is a non-Hermitian matrix.

Both errors ε_X and ε_P tend to zero in the limit $X \rightarrow X_{\text{exact}}$, but for a threshold in the range of a few percent, they behave differently. The condition upon ε_P is better tailored to the problem, and it saves unnecessary iterations in many cases. However, in a few cases, the condition upon ε_P fails, while the condition upon ε_X works. Therefore, we use a condition on ε_P in “quick and dirty” runs and a combined condition $\varepsilon_P < \varepsilon_{\text{tolerance}}$ and $\varepsilon_X < \varepsilon_{\text{tolerance}}$ for reliable results.

In general, iterative methods of the Krylov type involve only matrix–vector products $\mathbf{A}|z\rangle$. For an explicitly given matrix \mathbf{A} , the operation $|z\rangle \rightarrow \mathbf{A}|z\rangle$ requires $O(N^2)$ operations. Therefore, the whole iterative method will scale as $O(N^2 N_{\text{iter}})$, where N_{iter} is the number of iterations until convergence. This is better than direct methods when $N_{\text{iter}} \ll N$, because a matrix–matrix multiplication takes $O(N^3)$ operations.

To avoid matrix multiplications, the application of the matrix $\mathbf{A} = 1 - \chi^0(\omega) f_{\text{Hxc}}$ to a vector $|z\rangle$ is done sequentially by computing first $|z'\rangle = f_{\text{Hxc}}|z\rangle$ and then $\mathbf{A}|z\rangle = |z\rangle - \chi^0(\omega)|z'\rangle$. The kernel matrix f_{Hxc} is computed before the iterative procedure. Because it is frequency-independent, it can be easily stored and reused. By contrast, the response matrix $\chi^0(\omega)$ is frequency-dependent and computationally expensive, and its explicit construction should be avoided. Therefore, only matrix–vector products $\chi^0(\omega)|z\rangle$ will be computed as explained below without ever calculating the full response matrix $\chi^0(\omega)$.

4.2. Fast Application of the Kohn–Sham Response Matrix to Vectors. In previous papers,^{14,15} we have described an $O(N^2)$ construction of the entire response function $\chi_{\mu\nu}^0(\omega)$, but the prefactor was large. Paradoxically, the Krylov method presented above allows for a much faster computation of the absorption spectrum, although the cost of matrix–vector products $\chi^0(\omega)|z\rangle$ scales asymptotically as $O(N^3)$ (see below).

The starting point of our construction of the matrix–vector product $\chi^0(\omega)|z\rangle$ is the expression (eq 7) for the Kohn–Sham response matrix in the basis of dominant products:

$$\chi_{\mu\nu}^0(\omega) z^v = \sum_{abcd,E,F} (n_F - n_E) \frac{(X_a^E V_\mu^{ab} X_b^F)(X_c^F V_\nu^{cd} X_d^E)}{\omega - (E - F) + i\varepsilon} z^v$$

To compute this matrix–vector product efficiently, we decompose its calculation into a sequence of multiplications

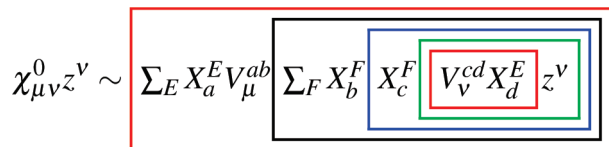


Figure 1. A sequence of operations to compute the matrix–vector product $\chi_{\mu\nu}^0 z^v$.

Table 1. Complexity and Memory Requirements during the Computation of the Matrix–Vector Product $\chi^0(\omega)z^a$

step	expression	complexity	memory	details of the complexity
1	$\alpha_v^{cE} = V_v^{cd} X_d^E$	$O(N^2)$	$O(N^2)$	$N_{\text{occ}} n_{\text{orb}}^2 N_{\text{prod}}$
2	$\beta^{cE} = \alpha_v^{cE} v^v$	$O(N^2)$	$O(N^2)$	$N_{\omega} N_{\text{occ}} n_{\text{orb}} N_{\text{prod}}$
3	$a^{FE} = X_c^F \beta^{cE}$	$O(N^3)$	$O(N^2)$	$N_{\omega} N_{\text{occ}} N_{\text{virt}} N_{\text{orb}}$
4	$\gamma_b^E = X_b^F a^{FE}$	$O(N^3)$	$O(N^2)$	$N_{\omega} N_{\text{occ}} N_{\text{virt}} N_{\text{orb}}$
5	$n_{\mu} = \sum_E X_a^E V_{\mu}^{ab} \gamma_b^E$	$O(N^2)$	$O(N)$	$N_{\omega} N_{\text{occ}} n_{\text{orb}} N_{\text{prod}}$

^a There are steps in the sequence with $O(N^2)$ and $O(N^3)$ complexity, where N is the number of atoms. N_{ω} denotes the number of frequencies for which the polarizability is computed, and the remaining symbols are explained in the text.

that minimizes the number of arithmetical operations by exploiting the sparsity of the vertex V_{μ}^{ab} . The sequence we chose is graphically represented in Figure 1. For clarity, the frequency-dependent denominator is omitted. Boxes represent the products to be performed at different steps. The innermost box contains a frequency-independent quantity that is also used in the last step. An algebraic representation of the computational steps is given in Table 1.

In the first stage of the algorithm, we compute an auxiliary object $\alpha_v^{cE} \equiv V_v^{cd} X_d^E$. The vertex V_v^{cd} is a sparse object by construction. Therefore, we will spend only $N_{\text{occ}} n_{\text{orb}}^2 N_{\text{prod}}$ operations, where N_{occ} is the number of occupied orbitals, N_{prod} is the number of products, and n_{orb} is the number of orbitals that belong to a pair of atoms. The auxiliary object α_v^{cE} is sparse and frequency-independent. Therefore, we store α_v^{cE} and reuse it in the fifth step of the matrix–vector product. The product $\beta^{cE} \equiv \alpha_v^{cE} v^v$ will cost $N_{\text{occ}} n_{\text{orb}} N_{\text{prod}}$ operations because each product index v “communicates” with the atomic orbital index c in one or two atoms. The matrix β^{cE} will be full; therefore the product $a^{FE} \equiv X_c^F \beta^{cE}$ will cost $N_{\text{occ}} N_{\text{virt}} N_{\text{orb}}$; i.e., this step has $O(N^3)$ complexity with N_{virt} being the number of unoccupied (virtual) states. The next step $\gamma_b^E \equiv \sum_F X_b^F a^{FE}$ also has $O(N^3)$ complexity with the same operation count. Finally, the sum $n_{\mu} \equiv \sum_E X_a^E V_{\mu}^{ab} \gamma_b^E$ takes only $O(N^2)$ operations because the vertex V_{μ}^{ab} is sparse. The sequence involves $O(N^3)$ operations, but due to prefactors, the run time is dominated by the $O(N^2)$ part of the sequence for molecules of up to 100 atoms.

4.3. Memory Requirements of the Algorithm. In our previous works,^{14,15} we computed TDDFT absorption spectra as a byproduct of the fast construction of the Kohn–Sham response function in $O(N^2 N_{\omega})$ operations. The memory requirement of the previous algorithm was very high because the use of FFT-based convolutions prohibited a frequency-by-frequency construction. The present work avoids the computation of the full-matrix response, and therefore it requires much less memory. In fact, one needs at most $O(N^2)$ storage in the application of the response function when

following the sequence of matrix operations outlined in subsection 4.2.

The interaction kernels will be computed before the iterative procedure (see section 5), and they also require $O(N^2)$ storage. This is more memory than in van Gisbergen’s iterative method.⁸ However, as we keep only the dominant products in our approach, N_{prod} is small compared to the number of the original orbital products, and today’s computers easily provide the memory needed for our algorithm. The storage of the interaction kernels in memory allows a full exploitation of the computational power of present day machines and implementation of an efficient iterative procedure. Its efficiency is essential because the dynamic polarizability is computed for many frequencies.

5. Calculation of the Interaction Kernels

In the adiabatic local-density approximation (ALDA), the interaction kernel

$$f_{\text{H}} + f_{\text{xc}} = \frac{\delta V_{\text{H}}}{\delta n} + \frac{\delta V_{\text{xc}}}{\delta n}$$

is a frequency-independent matrix. While the representation of the Hartree kernel is straightforward in our basis

$$f_{\text{H}}^{\mu\nu} = \iint d^3 r d^3 r' F^{\mu}(\mathbf{r}) \frac{1}{|\mathbf{r} - \mathbf{r}'|} F^{\nu}(\mathbf{r}') \quad (13)$$

the exchange–correlation kernel is not known explicitly and must be approximated. The locality of the LDA potential $V_{\text{xc}}(\mathbf{r}) = V_{\text{xc}}(n(\mathbf{r}))$ leads to a simple expression for the ALDA exchange–correlation kernel:

$$f_{\text{xc}}^{\mu\nu} = \int d^3 r F^{\mu}(\mathbf{r}) f_{\text{xc}}(\mathbf{r}) F^{\nu}(\mathbf{r}) \quad (14)$$

where $f_{\text{xc}}(\mathbf{r})$ is a (nonlinear) function of the density $n(\mathbf{r})$. In this section, we describe how the Hartree and ALDA exchange–correlation kernels are computed.

5.1. The Hartree Kernel. The basis functions $F^{\mu}(\mathbf{r})$ that appear in the interaction kernels 13 and 14 are built separately for each pair of atoms. They are either *local* or *bilocal* depending on whether the atoms of the pair coincide or not. While the local products are spherically symmetric functions, the bilocal products possess only axial symmetry. Because of their axial symmetry, the bilocal products have a particularly simple representation in a *rotated* coordinate system

$$F^{\mu}(\mathbf{r}) = \sum_{j=0}^{j_{\text{cutoff}}} F_j^{\mu}(r) Y_{jm}(\mathbf{R}\mathbf{r}'), \quad \mathbf{r}' = \mathbf{r} - \mathbf{c} \quad (15)$$

where the rotation \mathbf{R} and the center \mathbf{c} depend on the atom pair. In the rotated frame, the Z axis coincides with the line connecting the atoms in the pair. We use radial products $F_j^{\mu}(r)$ that are given on a logarithmic grid. The local products have a simpler, LCAO-like representation

$$F^{\mu}(\mathbf{r}) = F^{\mu}(|\mathbf{r} - \mathbf{c}|) Y_{jm}(\mathbf{r} - \mathbf{c}) \quad (16)$$

where the centers \mathbf{c} coincide with the center of the atom.

Using the algebra of angular momentum, we can get rid of the rotations \mathbf{R} in the bilocal basis functions (eq 15) and transform the kernel (eq 13) into a sum over two center Coulomb integrals $\text{Cb}(1, 2)$

$$\text{Cb}(1, 2) = \iint d^3r_1 d^3r_2 g_{l_1 m_1}(\mathbf{r}_1 - \mathbf{c}_1) |\mathbf{r}_1 - \mathbf{r}_2|^{-1} g_{l_2 m_2}(\mathbf{r}_2 - \mathbf{c}_2) \quad (17)$$

The elementary functions $g_{lm}(\mathbf{r}) = g_l(r) Y_{lm}(\mathbf{r})$ have a radial-angular decomposition similar to local products (eq 16). The Coulomb interaction (eq 17) becomes local in momentum space

$$\text{Cb}(1, 2) = \int d^3p g_{l_1 m_1}(\mathbf{p}) p^{-2} \exp(i\mathbf{p}(\mathbf{c}_1 - \mathbf{c}_2)) g_{l_2 m_2}(\mathbf{p}) \quad (18)$$

where the Fourier image of an orbital $g_{lm}(\mathbf{p})$ conserves its spherical symmetry

$$g_{lm}(\mathbf{p}) = i g_l(p) Y_{lm}(\mathbf{p}) \quad (19)$$

The radial part $g_l(p)$ is given by the Hankel transform of the original radial orbital in coordinate space:

$$g_l(p) = \sqrt{\frac{2}{\pi}} \int_0^\infty g_l(r) j_l(pr) r^2 dr \quad (20)$$

Inserting expression 19 into eq 18, expanding the plane wave $\exp(i\mathbf{p}(\mathbf{c}_1 - \mathbf{c}_2))$ in spherical harmonics, and using the algebra of angular momentum, we can reduce the integration in momentum space to a one-dimensional integral:

$$I_{l_1, l_2, l}(R) = \int_0^\infty g_{l_1}(p) j_l(pR) g_{l_2}(p) dp \quad (21)$$

where $R \equiv |\mathbf{c}_1 - \mathbf{c}_2|$.

We distinguish two cases in treating this integral according to whether the orbitals $g_{lm}(\mathbf{r})$ are overlapping or not. For overlapping orbitals, we compute the integral (eq 21) numerically using Talman's fast Hankel transform.²⁹ For nonoverlapping orbitals, one can compute the integral (21) exactly using a multipole expansion. To derive this expansion, we replace the functions $g_l(p)$ in the eq 21 with their Hankel transforms (20):

$$I_{l_1, l_2, l}(R) = \frac{2}{\pi} \int_0^\infty dr_1 g_{l_1}(r_1) r_1^2 \int_0^\infty dr_2 g_{l_2}(r_2) r_2^2 \int_0^\infty j_l(r_1) j_l(pR) j_l(r_2) dp \quad (22)$$

The integral over three spherical Bessel functions $I_{l_1, l_2, l}(r_1, r_2, R) \equiv \int_0^\infty j_{l_1}(r_1) j_l(pR) j_{l_2}(r_2) dp$ reduces to a simple separable expression³⁰ provided two conditions are satisfied, $R > r_1 + r_2$ (the basis functions do not overlap) and $0 \leq l \leq l_1 + l_2$ (the triangle relation for angular momentum). Under these conditions, we obtain

$$I_{l_1, l_2, l}(r_1, r_2, R) = \delta_{l, l_1 + l_2} \frac{\pi^{3/2} r_1^{l_1} r_2^{l_2}}{8 R^{l+1}} \frac{\Gamma(l + 1/2)}{\Gamma(l_1 + 3/2) \Gamma(l_2 + 3/2)} \quad (23)$$

Inserting this result into eq 22, we obtain an expression for the Coulomb interaction in terms of moments in closed form

$$\rho_l \equiv \int_0^\infty r^2 dr g_l(r) r^l \quad (24)$$

The moments (eq 24) can be computed and stored at the beginning of the calculation. Therefore, the calculation will consist of summing the angular-momentum coefficients, and this is clearly much faster than a direct numerical integration in eq 21.

The complexity of the near-field interactions will be proportional to the number of atoms N . The calculation of the multipoles (24) for the far-field interaction requires $O(N)$ mathematical operations. The remaining part of the far-field interactions (Wigner rotations) scales as N^2 with the number of atoms.

5.2. The ALDA Exchange-Correlation Kernel. Unlike the Hartree kernel, the exchange-correlation kernel (eq 14) is local, and we therefore compute it directly in coordinate space by numerical quadrature. The supports of the dominant products $F^u(\mathbf{r})$ and $F^v(\mathbf{r})$ in the integrand of eq 14 have, in general, the shape of a lens. Therefore, the support of the integrand will generally be an intersection of two lenses.

We found, however, that integration in spherical coordinates gives sufficiently accurate and quickly convergent results. This is because the important matrix elements involve neighboring dominant products and the support of the dominant products is large compared to the distance between their centers. Therefore, the integrands of the important matrix elements (eq 14) have nearly spherical support. This situation is illustrated by the cartoon in Figure 2.

For each pair of dominant products, we use spherical coordinates that are centered at the midpoint between them. We use Gauss–Legendre quadrature for integrating along the radial coordinate and Lebedev quadrature³¹ for integrating over the solid angle

$$F_{xc} = \sum_{i=1}^{N_r} G_i \sum_{j=1}^{N_\Omega} L_j f_{xc}(r_i, \Omega_j) \quad (25)$$

Here, G_i and r_i are weights and knots of Gauss–Legendre quadrature, and L_j and Ω_j are weights and knots of Lebedev quadrature. The number of points $N_r \times N_\Omega$ can be kept reasonably small (24×170 by default).

The most time-consuming part of the exchange-correlation kernel is the electronic density. We found that calculating the density using the density matrix

$$n(\mathbf{r}) = \sum_{ab} f^a(\mathbf{r}) D_{ab} f^b(\mathbf{r}), \quad \text{where } D_{ab} = \sum_{E < 0} X_a^E X_b^E \quad (26)$$

provides a linear scaling of the run time of f_{xc} with the number of atoms. However, a calculation of the density via molecular orbitals

$$n(\mathbf{r}) = \sum_{E < 0} \psi_E(\mathbf{r}) \psi_E(\mathbf{r}), \quad \text{where } \psi_E(\mathbf{r}) = \sum_a X_a^E f^a(\mathbf{r}) \quad (27)$$

is faster in many cases, although the run time scales quadratically with the number of atoms. In order to optimize the run time, rather than insist on linear scaling, we choose the calculational approach automatically depending on the

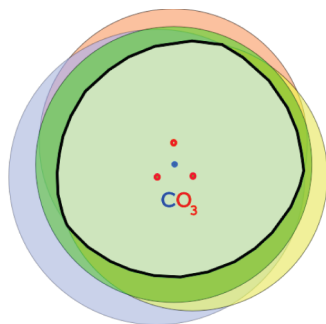


Figure 2. The spatial support of the integrand of the exchange correlation term (eq 14) depends on the support of its underlying atomic orbitals and on the geometry of the quadruplet of atoms under consideration. For neighboring atoms, the support of the orbitals is several times larger than their interatomic distances, and this results in a nearly spherical support of the integrand. The figure illustrates the case of (hypothetical) CO_3 .

Construction of the dominant products

for atom species do

Build local dominant products

!\$OMP PARALLEL DO

for atom pairs (a, b) do

Build bilocal dominant products

Computation of the interaction kernels f_H and f_{xc}

!\$OMP PARALLEL DO

for each couple of atom pairs $(a, b; c, d)$ do

Build block $(a, b; c, d)$ of kernels.

Computation of the dynamical polarizability

!\$OMP PARALLEL DO

for $\omega \in [\omega_{\text{begin}}, \omega_{\text{end}}]$ do

Solve for $|X\rangle$: $(1 - \chi^0(\omega)f_{Hxc})|X\rangle = \chi^0(\omega)|d\rangle$

Compute polarizability $P(\omega) = \langle d|X\rangle$

Figure 3. Skeleton of the algorithm. Its first (and computationally easiest) part is the construction of the dominant products. The second (and computationally most demanding) part is the construction of the interaction kernels. The third (and comparatively easy) part is the iterative calculation of the dynamic polarizabilities $P(\omega)$.

geometry of the molecule. For instance, in the case of a long polythiophene with 13 chains (see subsection 7.2), we use the $O(N)$ method, while in the other examples, it is better to use the $O(N^2)$ approach.

6. Parallelization of the Algorithm

The overall structure of our algorithm is given in Figure 3. First, the basis of dominant products is built. Then, the interaction kernels are computed, and finally both are used in the iterative procedure to compute the dynamic polarizability.

The individual components of the algorithm in Figure 3 suggest different parallelization strategies. The dominant products are built for each atom pair independently; therefore the corresponding code is parallelized over atom pairs. The

structure of the dominant products suggests a blockwise computation of the interaction kernels. These blocks are mutually independent; therefore we parallelize their construction. The dynamic polarizabilities are calculated independently for each frequency and are, therefore, parallelized over frequencies. Below, we go through the details of the algorithm and its hybrid OpenMP/MPI parallelization.

6.1. Multithread Parallelization. Modern computers are faster than previous generations of machines mainly due to their parallel design, as in the case of general-purpose multicore processors. For specially written programs, such a design allows several tasks or “threads” to run simultaneously. Fortunately, it is easy to write a multithreaded program using the current application programming interface OpenMP. Moreover, in OpenMP, data exchange between threads uses the common memory, and it is, therefore, faster than that on distributed-memory machines. For these reasons, our main emphasis here is on multithreaded (or shared-memory) parallelization. We use the OpenMP standard³² that allows for an efficient parallelization of all three sections of the algorithm in Figure 3.

6.1.1. Building the Basis of Dominant Products. The construction of local dominant products involves only the atomic species that occur in a molecule. Therefore, it is computationally cheap, and any parallelization would only slow down their construction.

For bilocal dominant products, the situation is different. The construction of bilocal dominant products is done for all atom pairs, the orbitals of which overlap. Because these pairs are independent of each other, we parallelize the loop over pairs with OpenMP directives.

The dominant products $F^u(\mathbf{r})$ and vertices V_μ^{ab} are stored in suitably chosen data structures to allow for effective use of memory. Due to the locality of our construction, the amount of memory spent in the storage of dominant products and vertices grows linearly with the number of atoms.

6.1.2. Construction of the Interaction Kernels. The interaction kernels 13 and 14 refer to a pair of dominant products $F^u(\mathbf{r})$ and $F^v(\mathbf{r})$. Because each of the dominant products in turn refers to a pair of atoms, the interaction kernel splits into blocks that are labeled by a quadruplet of atoms $(a, b; c, d)$. In practice, this is used to precompute and reuse some auxiliary quantities that belong to such a quadruplet. The block structure is schematically depicted in Figure 4. Generically, a block is rectangular, but when two pairs coincide, the block reduces to a triangle because of reflection symmetry.

The calculation of the interaction kernels is parallelized over blocks using dynamic scheduling. Because the computational load of each block is proportional to its area, we minimize the waiting time at the end of the loop by a descending sort of the blocks according to their area.

6.1.3. Computation of the Dynamic Polarizability. According to eq 10, the dynamic polarizability $P_{ik}(\omega)$ is computed independently for each frequency ω . Therefore, the loop over frequencies in the algorithm in Figure 3 is embarrassingly parallel.

In the parallelization over frequencies, we had to make thread-safe copies of module variables (working arrays in

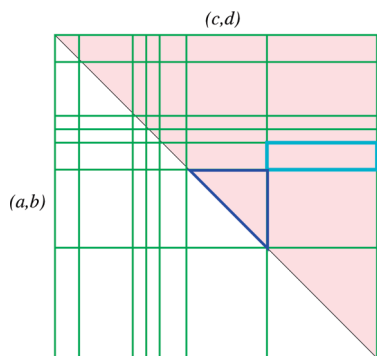


Figure 4. The block structure of the interaction kernels. The blocks are defined by two atom pairs (a, b) and (c, d) . The size of blocks is known from the construction of the dominant products. Because the entire interaction kernel $f_{\text{Hxc}}^{\text{AV}}$ is symmetric, only its upper triangular part is computed.

the GMRES solver) using the OpenMP directive threadprivate. This multiplies the memory requirement by the number of threads, but this poses no problem, because this part of the algorithm is not memory-intensive.

The number of iterations to reach convergence of the polarizability tensor (eq 10) varies with frequency and with the Cartesian components in an irregular way. To take this into account, we use a dynamic schedule with a single frequency and tensor component per thread. By treating tensor components on the same level as frequencies, we reduce the body of a loop by a factor of 3 provided only diagonal components are computed.

6.2. Hybrid MPI-Thread Parallelization. Most current supercomputers are organized as clusters of multicore nodes that are interconnected by a high-speed network. Although the number of cores has grown over the years, we still need several nodes for greater computational speed and to provide sufficient memory.

Our program has also been adapted to such distributed-memory parallel machines. We parallelize according to the Single Program Multiple Data (SPMD) paradigm with the message passing interface (MPI) to speed up only the computationally intensive parts of the algorithm—the construction of the bilocal products, the calculation of the kernels f_{H} and f_{xc} , and the iterative procedure. Moreover, each MPI process uses the multithread parallelism described above.

6.2.1. Parallelization of the Basis of Dominant Products. Only bilocal dominant products must be constructed in parallel. As described in subsection 6.1.1, the construction is naturally parallel in terms of atom pairs. Therefore, we distribute the atom pairs prior to computation and gather data after the computation in order to duplicate basis functions on each MPI process.

6.2.2. Parallelization of the Interaction Kernels. As explained in subsection 6.1.2, the interaction kernel depends on quadruplets of atoms, and it would appear natural to parallelize their construction over these quadruplets. We found it advantageous, however, to slice the interaction matrix into vertical bands that belong to one or more atom pairs and process them on different nodes. After the computation, the complete matrix of interactions is reconstituted on each node.

The optimal size of each vertical band is determined by an estimate of the work load prior to the computation. Since the Hartree and the exchange-correlation kernel differ in their properties (such as locality), their matrices are distributed differently. In both cases, however, the total work load is the sum of the work loads of its constituent quadruplets.

In the case of the Hartree kernel (eq 13), the work load of a block depends on its size and on whether its atom pairs are local or bilocal. While the local products are of the simple LCAO type (eq 16), the bilocal products (eq 15) contain additional summations. We found the following robust estimator of the workload of a block of the Hartree kernel:

$$\text{Workload}(\text{Hartree}) = \text{Size_Of_Block} \cdot (j_{\text{cutoff}} \cdot \Theta(a \neq b) + 1) \cdot (j_{\text{cutoff}} \cdot \Theta(c \neq d) + 1) \quad (28)$$

where $\Theta(a \neq b)$ is equal to 1 for bilocal atom pairs and 0 otherwise; j_{cutoff} is the largest angular momentum in the expansion (eq 15). By default, its upper limit is set to 7.

In the exchange-correlation kernel (eq 14), the domain of integration is the intersection of two lens-like regions. Because there is no simple analytical expression for such a volume, we count the number of integration points within the support for each block and estimate the work load as proportional to the number of points. Rather surprisingly, the run time is independent of the dimension of the block (this is due to precomputing the values of the dominant products).

Because the kernel f_{Hxc} is frequency-independent, it is computed at the beginning of the iterative procedure and duplicated on all the MPI nodes. We found the time for gathering the matrix small compared to the time of its computation.

6.2.3. Parallelization of the Iterative Procedure. As mentioned previously in subsection 6.1.3, the iterative calculation of the polarizability tensor $P_{ik}(\omega)$ is naturally parallel both in its frequency and in its Cartesian tensor indices. Therefore, this part of the algorithm is parallelized over both frequency and tensor indices. However, the workload for each composite index is difficult to predict. To achieve a balanced workload on the average, we distribute this index cyclically over MPI nodes (“round-robin distribution”).

7. Results

In this section, we present different absorption spectra of large molecules to validate the method and the parallelization approach. We start in subsection 7.1 by comparing our absorption spectra with previous calculations by other authors and also with experimental data. Then, in subsection 7.2, we present the scaling of the run time with the number of atoms. In subsection 7.3, we examine the efficiency of the hybrid parallelization for a variety of molecules run on different machines. Finally, in subsection 7.4, we present a comparison of absorption spectra of fullerene C_{60} with its derivative PCBM that is often used in organic solar cells.

7.1. Fullerene C_{60} . We already tested the basis of dominant products in our previous works,^{13–15} where the absorption spectra of methane, benzene, indigo blue, and

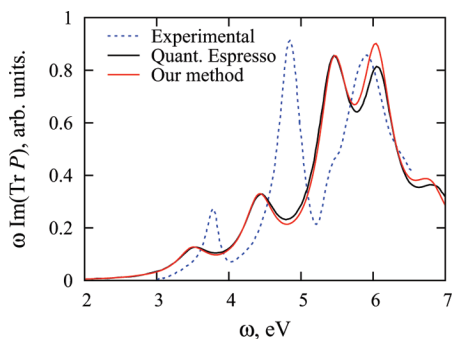


Figure 5. Comparison of the low-frequency absorption spectra of C_{60} fullerene. We see good agreement between the two theoretical predictions, while the experimental curve is shifted by 0.2–0.3 eV.

buckminster fullerene C_{60} were studied. The new element in the present paper is the use of an iterative method without constructing the full Kohn–Sham response function. We now test this method on buckminster fullerene.

Buckminster fullerene C_{60} is of considerable interest in materials science. Among other applications, it is used as an electron acceptor in organic solar cells.³³ Here, we compare our results with spectra from the Quantum Espresso package and with experimental results.

The absorption spectrum of fullerene C_{60} is shown in Figure 5 where we compare our results with the calculation by Rocca et al.¹¹ and with experimental results.³⁴ One can see a good agreement between the theoretical spectra, while both theoretical predictions deviate from the experimental data by 0.2–0.3 eV. The shift in the spectrum might be due to the solvent in the experimental setup or due to the inadequacy of the simplest LDA functional for this large molecule. We used DFT data from the SIESTA package,³⁵ where the pseudopotentials of the Troullier–Martin type and the LDA functional by Perdew and Zunger³⁶ are applied. A double- ζ polarized basis set has been used, and the broadening of levels has been set to 0.019 Ry. Our program spent a total of about 62.5 min on this calculation, of which 2098 s were spent on the Coulomb kernel, 1085 s on the exchange-

correlation kernel, and 500 s in the iterative procedure, i.e., approximately 2.27 s per frequency. The convergence parameter for the polarizability was set to 1%. With this convergence parameter, the dimension of the Krylov space was varying from 7 to 12, with an average of 8, while the dimension of the dominant product space was 8700. In this test, we used one thread on a machine with two Intel quad core Nehalem CPUs at 2.93 GHz, with an 8 MB cache, and 48 GB of DDR3 RAM, and consuming no more than 2.3% of RAM (1.2% during the iterative procedure).

7.2. Polythiophene Chains (Complexity of the Method). In sections 4 and 5, we discussed the complexity scaling of different parts of the algorithm theoretically. In this subsection, we measure the dependence of run time on the number of atoms N in polythiophene chains of different lengths. We shall see that the run time scaling follows the theoretical predictions for the complexity.

Sulfur-containing molecules are widely used in organic electronics.^{33,37,38} In this work, we study pure polythiophene chains of 3–13 repeating units. The geometry of the longest polythiophene we considered is shown in Figure 6. Our calculations suggest (ignoring the excitonic character of these molecules) that the HOMO–LUMO energy difference decreases, while the absorption increases, with chain length. The calculated absorption spectra are collected in Figure 7.

We now use the calculations on polythiophene spectra in order to study the run time scaling with the number of atoms of different parts of our algorithm. Their scaling behavior will be described in terms of approximate scaling exponents. The run times for a few chains are collected in Table 2 for a machine of four AMD Dual-Core Opteron CPUs at 2.6 GHz with an 8 MB cache, and 32 GB of DDR2 RAM, running sequentially.

The application of the noninteracting response to a vector consists of N^2 and N^3 parts (see subsection 4.2). The total time for the product $\chi^0 z$ is collected in the third column of Table 2, while the run time of the N^3 part is collected in the fourth column. Using the run times, one can compute

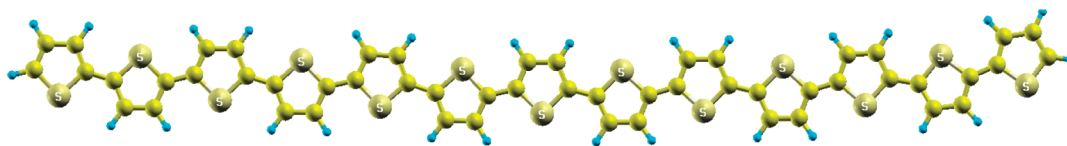


Figure 6. The geometry of the longest polythiophene chain we considered.

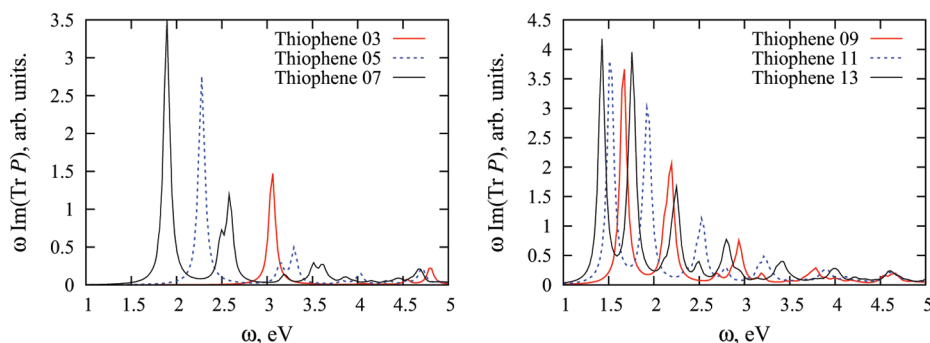
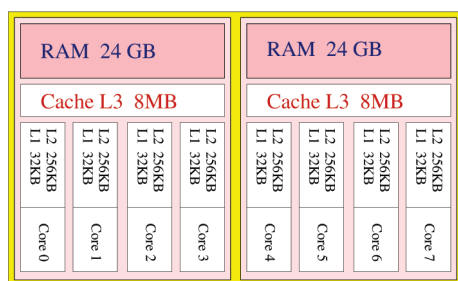


Figure 7. Comparison of low-frequency spectra for several polythiophene chains.

Table 2. Run Time in Different Parts of the Algorithm as a Function of the Number of Atoms N in the Polythiophene Chain^a

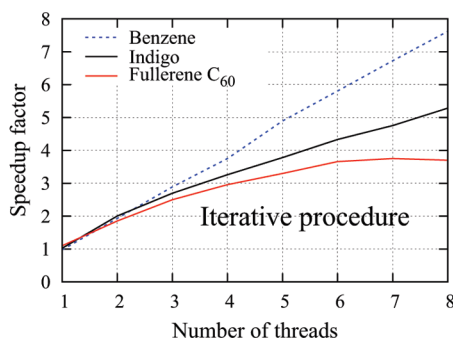
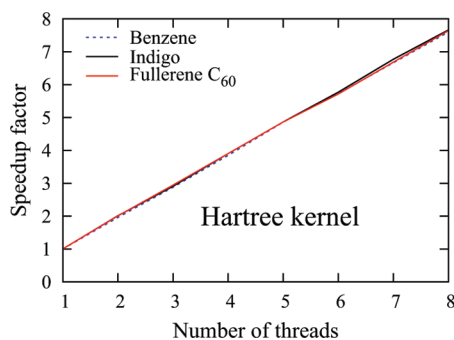
N	units	χ^0z , s	N^3 part in χ^0z , s	f_H , s	f_{xc} , s
23	3	1.51×10^{-2}	8.20×10^{-3}	157	173
37	5	4.42×10^{-2}	2.52×10^{-2}	417	300
51	7	9.29×10^{-2}	5.58×10^{-2}	807	428
65	9	0.166	0.104	1324	561
79	11	0.260	0.169	1977	694
93	13	0.382	0.255	2767	822

^a The third column gives the total time for the matrix–vector product (χ^0z), while an N^3 part that arises in the construction of χ^0z is given in the fourth column. The fifth and sixth columns display the run times of the interaction kernels f_H and f_{xc} , respectively.

**Figure 8.** Memory/cache/core structure of Intel Xeon X5550 Nehalem machine. Two quad core nodes have fast access to one of the memory banks, while the internode communication is slower.

exponents x and x_3 for their corresponding scaling laws N^x and N^{x_3} . The exponents x and x_3 vary in the range of $x = 2.31$ – 2.36 and $x_3 = 2.49$ – 2.53 , respectively. The run time of the Hartree kernel (fifth column) shows a scaling exponent in the range $x_H = 2.05$ – 2.06 , while the run time in the exchange–correlation kernel (sixth column) scales almost linearly $x_{xc} = 1.04$ – 1.12 . Therefore, the measured exponents are close to the predicted exponents $x = 3$, $x_3 = 3$, $x_H = 2$, and $x_{xc} = 1$.

The calculation of the Hartree kernel f_H via multipoles, as explained in subsection 5.1, improves the run time in the case of large molecules but could not improve the run time scaling of the Coulomb interaction. In fact, the Hartree kernel f_H is a nonlocal quantity, and the rotations involved in the bilocal dominant products contribute a substantial part to the run time.

**Figure 9.** (Left panel) Speedup factor in the Hartree kernel for molecules of different size. The speedup in the exchange–correlation kernel and in the generation of dominant products is similar to the left panel. The right panel shows the speedup factor in the iterative procedure. We observe that the speedup decreases with the size of the molecule.

The scaling of the run time for the entire calculation of molecular spectra will also vary with the parameters of the calculation. Obviously, for a small number of frequencies, the scaling of total run time will be determined by the Hartree and exchange–correlation kernels. However, if the number of frequencies is large, then the application of the noninteracting response χ^0 on a vector will dominate the run time.

7.3. Quality of the Parallelization. Our parallel Fortran 90 code is adapted to current parallel architectures (see section 6). In this subsection, we evaluate the quality of the hybrid parallelization by testing our approach on three machines of different architecture. Two machines belong to shared-memory multicore architectures with nonuniform memory access, while the third machine is a cluster with 50 multicore nodes interconnected by an Infiniband network.

The program was compiled using Intel’s Fortran compiler and linked against Intel’s Math Kernel Library for BLAS, LAPACK, and Fast Fourier Transform libraries.

7.3.1. Multithread Parallelization. We consider parallelization on shared-memory machines as more important than on distributed-memory machines. Therefore, the speedup of our code is first tested on two shared-memory machines with Non-Uniform Memory Architecture (NUMA). The first machine has two quad core Nehalem 5500 CPUs (see Figure 8), while the second machine has 48 dual core Xeon CPUs (see Figure 10).

The speedup on the Nehalem machine is shown in Figure 9 for three molecules of different size: benzene, indigo, and fullerene. The speedup in both kernels and in the generation of bilocal products is good for all molecules; therefore we plot only the speedup for the Hartree kernel on the left panel. By contrast, the iterative procedure exhibits a lower speedup (as shown on the right panel in Figure 9). This is due to the high memory–bandwidth requirement of the iterative procedure. The high memory bandwidth is clearly revealed in two distinct ways of using the same number of cores in a test calculation, see subsection 7.3.2.

In spite of the loss of speedup for larger molecules on the Nehalem machine, we tested our implementation also on a parallel processor with a very large number of cores (see Figure 10). The speedup is shown in Figure 11.

Our results show a satisfactory speedup in both interaction kernels, while the iterative procedure again shows a poorer

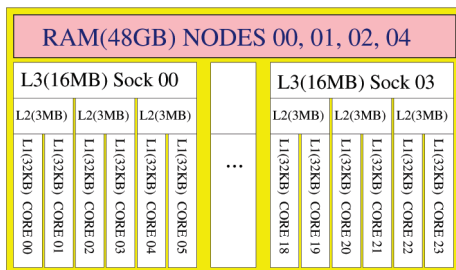


Figure 10. Memory/cache/core structure of Xeon-96 machine. A total of 48 dual core Xeon CPUs are connected to four memory banks. Although every core can address the whole memory space, the internode communication is slower.

performance due to its high memory-bandwidth requirements, which cannot be satisfied in this NUMA architecture.

7.3.2. Hybrid MPI-Thread Parallelization. Prior to large scale computations with many Nehalem (see Figure 8) nodes, we performed test runs on one node to find an optimal OpenMP/MPI splitting. Four calculations were done on the Nehalem machine using all eight cores of the machine, but differing in the OpenMP/MPI splitting. The results are collected in the Table 3. We observed a small workload unbalance of 8% in case of the Hartree kernel and an even smaller unbalance of 5% in the case of the exchange-correlation kernel. The iterative procedure with the “round robin distribution” of frequencies over the nodes shows an even lower workload unbalance of 2%. The best run time is achieved in a 2/4 hybrid parallelization, i.e., running two processes with four cores each. This optimal configuration reflects the structure of the machine, where each thread shares the same L3 cache. The node has two processors of four cores each and a relatively slower memory access between the processors. This weak internode communication results in an appreciable penalty in an OpenMP-only run (1/8 configuration), because the iterative procedure reads a rather large amount of data ($V_{\mu}^{ab} X_b^E$) twice during the application of the response function χ_0 (see subsection 4.2).

The high memory-bandwidth requirement is clearly revealed in two distinct ways of using the same number of cores (eight cores), either using all cores on one node or distributing them over two nodes. In the latter case, the memory-bandwidth is higher and the iterative procedure runs considerably faster (92 s versus 137 s in the case of fullerene C_{60}).

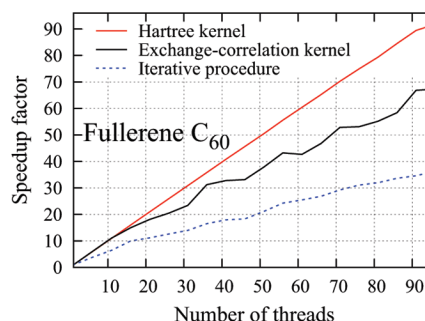
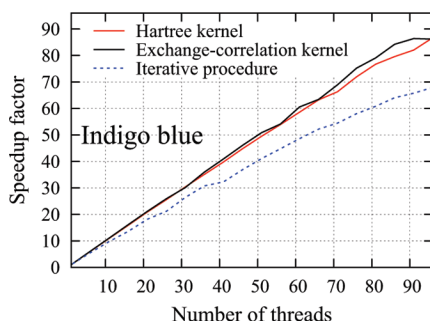


Figure 11. Speedup on a heavily parallel Xeon-96 machine. The results are satisfactory for the interaction kernels, while the memory-bandwidth requirements of the iterative procedure hamper the parallel performance in the case of the larger fullerene C_{60} molecule.

Table 3. Run Time (in seconds) in a Hybrid MPI/OpenMP Parallelization for Fullerene C_{60} ^a

proc/thr	domi. prod.	f_H	f_{xc}	iterative proc.	total
1/8	7.3	271	142	145	571
2/4	6.9 (6.9)	273 (267)	142 (141)	109 (108)	538
4/2	6.8 (6.8)	274 (264)	142 (141)	122 (112)	544
8/1	6.8 (6.8)	274 (257)	143 (140)	134 (120)	570

^a In the parentheses, the smallest run time between the nodes is stated in order to estimate the MPI work load disbalance.

We used the above optimal 2/4 hybrid configuration in a massively parallel calculation on the chlorophyll-a molecule. The speedup due to hybrid OpenMP-MPI parallelization is shown in Figure 12. In this computation, we used up to 50 nodes of recent generation Nehalem machines. According to the previous experiment on fullerene C_{60} , we started two processes per node, each process running with four threads. The two processes were placed on sockets. One can see that the iterative procedure shows the best speedup among other parts of the code, while total run time is governed mainly by the calculation of the exchange-correlation kernel. The absolute run times (including communication time) in the first calculation with one node (two processes) were: total, 4003 s; for the exchange-correlation kernel, 1147 s; for the Hartree kernel, 1247 s; for the iterative procedure, 1574 s; and for the bilocal vertex, 11.8 s. The speedup in the bilocal vertex reaches a maximum at 10 nodes because of increasing communication time.

The starting geometry of the molecule was taken from Sundholm’s supplementary data.³⁹ The geometry was further relaxed in the SIESTA package³⁵ using Broyden’s algorithm until the remaining force was less than 0.04 Ry/Å. The relaxed geometry is shown in Figure 13. In order to achieve this (default) criterion, we had to use a finer internal mesh with a MeshCutoff of 185 Ry. The default DZP basis was used, but to achieve convergence, we used orbitals that are more extended in space than SIESTA’s default orbitals. The spatial extension is governed by the parameter PAO. EnergyShift that was set to 0.002 Ry in the present calculation. The spectrum of the chlorophyll-a molecule is seen in Figure 14. Like in the case of fullerene C_{60} , there is excellent agreement between theoretical results that however differ from the experimental data.⁴⁰ The low frequency spectrum consists of two bands at 635 and 450 nm. Both bands are due to transitions in the porphyrin. According to our calculation, the first A band consists of two transitions

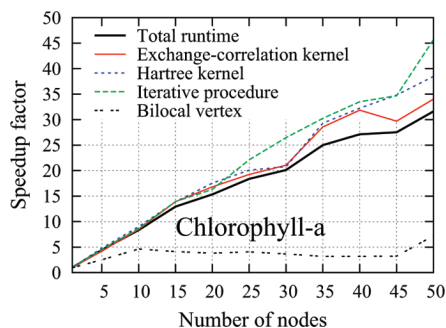


Figure 12. Speedup due to hybrid OpenMP/MPI parallelization for chlorophyll-a. The job was run on up to 50 nodes with two processes per node. The code shows a linear speedup on up to 15 nodes (30 processes, 120 cores). Further increase of the number of nodes results in a steady acceleration of the whole program.

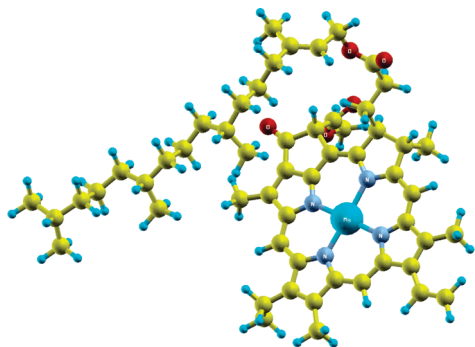


Figure 13. The relaxed geometry of the chlorophyll-a molecule obtained with the SIESTA package using a DZP basis set.

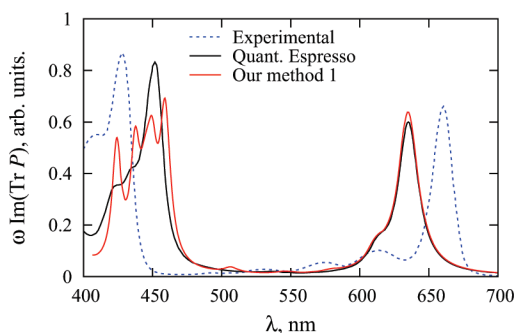


Figure 14. Low frequency absorption spectrum of chlorophyll-a.

between HOMO–LUMO and HOMO–1–LUMO, while the second (so-called Soret) band consist of six transitions to LUMO and LUMO+1 states.

7.4. Fullerene C₆₀ versus PCBM. Fullerenes are often modified in order to tune their absorption spectra or their transport properties.^{33,41,42} In this work, we compute the absorption spectra of [6,6]-phenyl C₆₁ butyric acid methyl ester (PCBM) and compare it with the spectrum of pure fullerene C₆₀. We use the same parameters as in the case of C₆₀ in subsection 7.1. A relaxed geometry of PCBM was obtained using the SIESTA package³⁵ and using its default convergence criterion (maximal force less than 0.04 eV/Å). Figure 15 shows the relaxed geometry. The absorption spectrum of PCBM is shown in Figures 16 and 17.

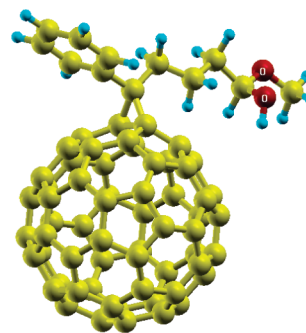


Figure 15. Geometry of PCBM. Relaxation is done in the SIESTA package with Broyden's algorithm.

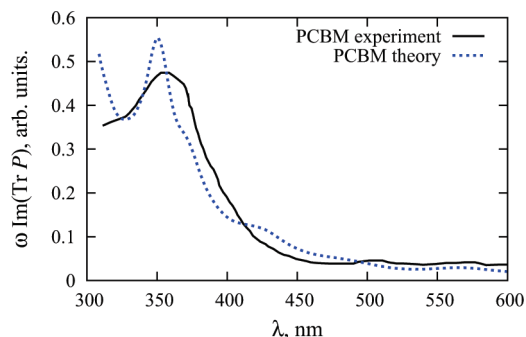


Figure 16. Comparison of the low-frequency spectra for PCBM with experimental data.

Figure 16 shows a comparison of our calculation with recent experimental results.⁴² We can see that our results have similar features as the experimental data: the maxima at 350 nm agree well with a broad experimental resonance at 355 nm, and a substantial background at a longer wavelength is present both in the calculated and in the experimental spectrum. In this calculation, we set the damping constant $\varepsilon = 0.08$ Ry and compute the spectrum in the range where experimental data are available. However, in order to better understand the difference introduced by the functional group, we compute the spectra in a broader range of energies with a smaller value of the damping constant $\varepsilon = 0.003$ Ry. The result is shown in Figure 17.

One can see on the left panel of Figure 17 that PCBM absorbs much stronger in the visible range. This is a consequence of symmetry breaking and indicates a modified HOMO–LUMO gap. On the right panel of the figure, one can recognize the main difference between pure and modified fullerene. The high spatial symmetry of pure fullerene leads to a degeneracy of the electronic transitions, and several transitions contribute to the same resonance. The symmetry is broken in the case of PCBM. The degeneracy is lifted, and the spectral weight is spread out.

8. Conclusion and Outlook

In this paper, we have described a new iterative algorithm for computing molecular spectra. The method has two key ingredients. One is a previously constructed local basis in the space of products of atomic LCAO orbitals. The second is the computation of the density response not in the entire space of products, but in an appropriate Krylov subspace.

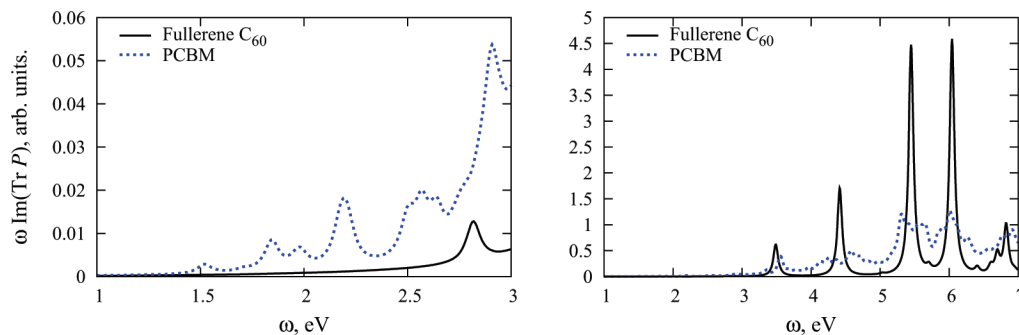


Figure 17. Comparison of the low-frequency spectra C_{60} versus PCBM.

The speed of our code is roughly comparable to TDDFT codes in commercially available software. The reader can judge the speed from the absolute timings we presented above. In general, we expect our method to be faster than the solution of Casida's equation for systems with dense spectra, when the target range contains many allowed transitions.

The algorithm was parallelized and was shown to be suitable for treating molecules of more than 100 atoms on large current heterogeneous architectures using the OpenMP/MPI framework.

Our approach leaves plenty of room for further improvements both in the method and in the algorithm. For example, we did not consider reducing the dimension of the space in which the response function acts, but such a reduction is feasible.

Also, we are working on an adaptive procedure to obtain good spectra with fewer frequency points and to compute the position of the poles and strength of their residues. A reduction in the number of frequencies will allow avoidance of the full calculation of the interaction kernels, replacing them by matrix–vector multiplications. There exist fast multipole methods⁴³ for computing fast matrix–vector products needed in the Hartree interaction.

Moreover, for large molecules, our embarrassingly parallel approach to compute spectra induces a memory-bandwidth bottleneck. To avoid it, one may parallelize the frequency loop using MPI and parallelize the matrix–vector operations using OpenMP.

More generally, because we do not use Casida's equations, the methods developed here should be useful beyond the TDDFT approach, for instance, in the context of Hedin's GW approximation⁴⁴ where Casida's approach is no longer available.

Our framework appears suitable for computing the Hartree–Fock exchange energy and for treating hybrid functionals and their response. In fact, the calculation of the standard Hartree–Fock integrals reduces to the Hartree kernel $f_{\text{H}}^{\mu\nu}$ presented in this paper:

$$\int f d^3r d^3r' f^a(\mathbf{r}) f^b(\mathbf{r}') |\mathbf{r} - \mathbf{r}'|^{-1} f^c(\mathbf{r}) f^d(\mathbf{r}') = \sum_{\mu, \nu} V_{\mu}^{ab} f_{\text{H}}^{\mu\nu} V_{\nu}^{cd}$$

The calculation takes $O(N^2)$ operations because the vertices V_{μ}^{ab} are sparse by construction. Introduction of a screened Coulomb potential⁴⁵ would give linear scaling. The calcula-

tion of interaction kernels for hybrid functionals in our framework must be reserved to future work, however.

Acknowledgment. We are indebted to Gustavo Scuseria for calling our attention to the existence of iterative methods in TDDFT and to Stan van Gisbergen for correspondence on the iterative method implemented in the Amsterdam Density Functional package (ADF). It is our special pleasure to thank James Talman (University of Western Ontario, Canada) for contributing two crucial computer codes to this project. We thank Luc Giraud (HiePACS, Toulouse) for discussions on the GMRES algorithm and our colleagues Aurelian Esnard and Abdou Guermouch (University of Bordeaux) for technical advice. We acknowledge useful correspondence on the SIESTA code by Daniel Sanchez-Portal (DIPC, San Sebastian) and also by Andrei Postnikov (Verlaine University, Metz). Advice by our colleagues in the ANR project CIS 2007 “NOSSI”, especially Ross Brown and Isabelle Baraille (IPREM, Pau), is gratefully acknowledged. We also thank Uwe Huniak (Karlsruhe, Turbomole) for kindly supplying benchmarks of the TURBOMOLE package for comparison. The results and benchmarks of this paper were obtained using the PlaFRIM experimental testbed of the INRIA PlaFRIM development project funded by LABRI, IMB, Conseil Régional d'Aquitaine, FeDER, Université de Bordeaux and CNRS (see <https://plafrim.bordeaux.inria.fr/>). This work was performed in partial fulfillment of ANR CIS-007-05 “NOSSI” project.

References

- (1) Gross, E. K. U.; Burke, K. Basics. In *Time-Dependent Density Functional Theory*, 1st ed.; Marques, M. A. L., Ullrich, C. A., Nogueira, F., Rubio, A., Burke, K., Gross, E. K. U., Eds.; Springer: Berlin, Germany, 2008; pp 243–353 (part IV), p 231, p 9.
- (2) Perdew, J. P.; Kurth, S. Density Functionals for Non-relativistic Coulomb Systems in the New Century. In *A Primer in Density Functional Theory*, 1st ed.; Fiolhais, C., Nogueira, F., Marques, M. A. L., Eds.; Springer: Berlin, Germany, 2003; pp 1–55.
- (3) Runge, E.; Gross, E. K. U. Density-Functional Theory for Time-Dependent Systems. *Phys. Rev. Lett.* **1984**, *52*, 997.
- (4) Marques, M. A. L.; Gross, E. K. U. Time-Dependent Density-Functional Theory. *Annu. Rev. Phys. Chem.* **2004**, *55*, 427.
- (5) Petersilka, M.; Gossmann, U. J.; Gross, E. K. U. Excitation Energies from Time-Dependent Density-Functional Theory. *Phys. Rev. Lett.* **1996**, *76*, 1212.

- (6) Casida, M. E. Time-Dependent Density-Functional Response Theory for Molecules. In *Recent Advances in Density Functional Theory*, 1st ed.; Chong, D. P., Ed.; World Scientific: Singapore, 1995; p 155.
- (7) Casida, M. E. Time-dependent density-functional theory for molecules and molecular solids. *THEOCHEM* **2009**, *914*, 3.
- (8) van Gisbergen, S. J. A.; Fonseca Guerra, C.; Baerends, E. J. Towards Excitation Energies and (Hyper)polarizability Calculations of Large Molecules. Application of Parallelization and Linear Scaling Techniques to Time-Dependent Density Functional Response Theory. *J. Comput. Chem.* **2000**, *21*, 1511.
- (9) de Boeij, P. L. Solution of the Linear-Response Equations in a Basis Set In *Time-Dependent Density Functional Theory*, 1st ed.; Marques, M. A. L., Ullrich, C. A., Nogueira, F., Rubio, A., Burke, K., Gross, E. K. U., Eds.; Springer: Berlin, Germany, 2008; pp 211–215.
- (10) Jensen, J.; Autschbach, J.; Schatz, G. C. Finite lifetime effects on the polarizability within time-dependent density-functional theory. *J. Chem. Phys.* **2005**, *122*, 224115. Seth, M.; Ziegler, T.; Banerjee, A.; Autschbach, J.; van Gisbergen, S. J. A.; Baerends, E. J. Calculation of the *A* term of magnetic circular dichroism based on time dependent-density functional theory I. Formulation and implementation. *J. Chem. Phys.* **2004**, *120*, 10942.
- (11) Rocca, D.; Gebauer, R.; Saad, Y.; Baroni, S. Turbo charging time-dependent density-functional theory with Lanczos chains. Theoretical and experimental results for C₆₀ and chlorophyll-a. *J. Chem. Phys.* **2008**, *128*, 154105.
- (12) Rocca, D. Time-dependent density functional perturbation theory. Ph.D. Thesis, Scuola Internazionale Superiore di Studi Avanzati, Trieste, Italy, 2004.
- (13) Foerster, D. Elimination, in electronic structure calculations, of redundant orbital products. *J. Chem. Phys.* **2008**, *128*, 034108.
- (14) Foerster, D.; Koval, P. On the Kohn-Sham density response in a localized basis set. *J. Chem. Phys.* **2009**, *131*, 044103.
- (15) Koval, P.; Foerster, D.; Coulaud, O. Fast construction of the Kohn-Sham response function for molecules. *Phys. Status Solidi B* **2010**, *247*, 1841.
- (16) Fetter, A. L.; Walecka, J. D. *Quantum Theory of Many-Particle Systems*, 1st ed.; McGraw-Hill: New York, 1971; p 214.
- (17) Beebe, N. H. F.; Linderberg, J. Simplifications in the generation and transformation of two-electron integrals in molecular calculations. *Int. J. Quantum Chem.* **1977**, *7*, 683.
- (18) Boys, S. F.; Shavitt, I. A. Fundamental Calculation of the Energy Surface for the System of Three Hydrogen Atoms. Technical Report WIS-AF-13; University of Wisconsin Naval Research Laboratory: Madison, WI, 1959.
- (19) Skylaris, C.-K.; Gagliardi, L.; Handy, N. C.; Ioannou, A. G.; Spencer, S.; Willetts, A. On the resolution of identity Coulomb energy approximation in density functional theory. *THEOCHEM* **2000**, *501–502*, 229.
- (20) Baerends, E. J.; Ellis, D. E.; Ros, P. Self-consistent molecular Hartree-Fock-Slater calculations I. The computational procedure. *Chem. Phys.* **1973**, *2*, 41.
- (21) Te Velde, G.; Bickelhaupt, F. M.; Baerends, E. J.; Fonseca Guerra, C.; van Gisbergen, S. J. A.; Snijders, J. G.; Ziegler, T. Chemistry with ADF. *J. Comput. Chem.* **2001**, *22*, 931.
- (22) Vysotskiy, V. P.; Cederbaum, L. S. On the Cholesky Decomposition for electron propagator methods: General aspects and application on C60. 2009, arXiv: physics/0912.1459. arXiv.org ePrint archive. <http://arxiv.org/abs/0912.1459> (accessed Jun 29, 2010).
- (23) Aryasetiawan, F.; Gunnarsson, O. Product-basis method for calculating dielectric matrices. *Phys. Rev. B* **1994**, *49*, 16214.
- (24) Larrue, U. *Etude de la Densité spectrale d'une métrique associée à l'équation de Schrödinger pour l'hydrogène*. Unpublished, Bordeaux, 2008. This study showed an asymptotically uniform density of eigenvalues, on a logarithmic scale, for the metric in the case of the hydrogen atom.
- (25) Saad, Y. *Iterative Methods for Sparse Linear Systems*, 2nd ed; SIAM: Philadelphia, PA: 2000; p 158.
- (26) Frayssé, V.; Giraud, L.; Gratton, S.; Langou, J. Algorithm 842: A set of GMRES routines for real and complex arithmetics on high performance computers. *ACM Trans. Math. Softw.* **2005**, *31*, 228. <http://doi.acm.org/10.1145/1067967.1067970> (accessed Jun 29, 2010).
- (27) Ipsen, I. C. F.; Meyer, C. D. The Idea behind Krylov Methods. *Am. Math. Monthly* **1998**, *105*, 889.
- (28) Barrett, R.; Berry, M. W.; Chan, T. F.; Demmel, J.; Donato, J.; Dongarra, J.; Eijkhout, V.; Pozo, R.; Romine, Ch.; van der Vorst, H. *Templates for the Solution of Linear Systems: Building Blocks for Iterative Methods*, 2nd ed; SIAM: Philadelphia, PA, 1993; p 51.
- (29) Talman, J. D. LSFSTR: a subroutine for calculating spherical Bessel transforms. *Comput. Phys. Commun.* **1983**, *30*, 93. Talman, J. D. NumSBT: A subroutine for calculating spherical Bessel transforms numerically. *Comput. Phys. Commun.* **2009**, *180*, 332.
- (30) Gradsteyn, I. S.; Ryzhik, I. M. *Tables of Integrals, Series and Products*, 5th ed; Academic Press: New York, 1980; formula 6.578/4.
- (31) Lebedev, V. I. *Russ. Acad. Sci. Dokl. Math.* **1995**, *50*, 283. <http://www.ccl.net/ccs/software/SOURCES/FORTRAN/Lebedev-Laikov-Grids/> (accessed Jun 30, 2010).
- (32) The OpenMP API specification for parallel programming. <http://openmp.org/wp/> (accessed Jun 30, 2010).
- (33) Brabec, Ch. J.; Sariciftci, N. S.; Hummelen, J. C. Plastic Solar Cells. *Adv. Funct. Mater.* **2001**, *11*, 15.
- (34) Bauernschmit, R.; Ahlrichs, R.; Hennrich, F. H.; Kappes, M. M. Experiment versus Time Dependent Density Functional Theory Prediction of Fullerene Electronic Absorption. *J. Am. Chem. Soc.* **1998**, *120*, 5052.
- (35) Ordejón, P.; Artacho, E.; Soler, J. M. Self-consistent order-*N* density-functional calculations for very large systems. *Phys. Rev. B* **1996**, *53*, R10441. Soler, J. M.; Artacho, E.; Gale, J. D.; García, A.; Junquera, J.; Ordejón, P.; Sánchez-Portal, D. The SIESTA method for ab initio order-*N* materials simulation. *J. Phys.: Condens. Matter* **2002**, *14*, 2745. <http://www.icmab.es/siesta/> (accessed Jun 30, 2010). We used different branches of SIESTA version 3 to perform calculations in this work.
- (36) Perdew, J. P.; Zunger, A. Self-interaction correction to density-functional approximations for many-electron systems. *Phys. Rev. B* **1981**, *23*, 5048.
- (37) Liang, F.; Lu, J.; Ding, J.; Movileanu, R.; Tao, Ye. Design and Synthesis of Alternating Regioregular Oligothiophenes/Benzothiadiazole Copolymers for Organic Solar Cells. *Macromolecules* **2009**, *42*, 6107.

- (38) Gao, Y.; Martin, Th. P.; Thomas, A. K.; Grey, J. K. Resonance Raman Spectroscopic- and Photocurrent Imaging of Polythiophene/Fullerene Solar Cells. *J. Phys. Chem. Lett.* **2010**, *1*, 178.
- (39) Sundholm, D. Density functional theory calculations of the visible spectrum of chlorophyll a. *Chem. Phys. Lett.* **1999**, *302*, 480. Optimized geometry has been taken from supplementary at <http://www.chem.helsinki.fi/~sundholm/qc/chlorophylla/> (accessed Jun 30, 2010).
- (40) Du, H.; Fuh, R. C. A.; Li, J.; Corkan, L. A.; Lindsey, J. S. PhotochemCAD: A Computer-Aided Design and Research Tool in Photochemistry. *Photochem. Photobiol.* **1998**, *68*, 141. Database is available online at <http://omlc.ogi.edu/spectra/PhotochemCAD/html/index.html> (accessed Jun 30, 2010).
- (41) Mayer, A. C.; Scully, S. R.; Hardin, B. E.; Rowell, M. W.; McGehee, M. D. Polymer-based solar cells. *Mater. Today* **2007**, *10*, 28.
- (42) Suresh, P.; Balraju, P.; Sharma, G. D.; Mikroyannidis, J. A.; Stylianakis, M. M. Effect of the Incorporation of a Low-Band-Gap Small Molecule in a Conjugated Vinylene Copolymer: PCBM Blend for Organic, Photovoltaic Devices. *ACS Appl. Mater. Interfaces* **2009**, *1*, 1370.
- (43) Greengard, L.; Rokhlin, V. A fast algorithm for particle simulations. *J. Comput. Phys.* **1987**, *73*, 325. Greengard, L. Fast Algorithms for Classical Physics. *Science*. **1994**, *265*, 909.
- (44) Hedin, L.; Lundqvist, S. Effects of Electron-Electron and Electron-Phonon Interactions on the One-Electron States of Solids. In *Solid State Physics*, 1st ed; Academic Press: London, Great Britain, 1969; Vol. 23, pp 1–181.
- (45) Heyd, J.; Scuseria, G. E.; Ernzerhof, M. Hybrid functionals based on a screened Coulomb potential. *J. Chem. Phys.* **2003**, *118*, 8207.

CT100280X

Electric Field Gradients Calculated from Two-Component Hybrid Density Functional Theory Including Spin–Orbit Coupling

Fredy Aquino,[†] Niranjana Govind,[‡] and Jochen Autschbach^{*,†}

Department of Chemistry, State University of New York at Buffalo, Buffalo, New York 14260-3000, and William R. Wiley Environmental Molecular Sciences Laboratory, Pacific Northwest National Laboratory, 902 Battelle Blvd, P.O. Box 999, Mail Stop K8-91 Richland, Washington 99352

Received May 30, 2010

Abstract: An implementation of a four-component density corrected approach for calculations of nuclear electric field gradients (EFGs) in molecules based on the two-component relativistic zeroth-order regular approximation (ZORA) is reported. The program module, which is part of the NWChem package, allows for scalar and spin–orbit relativistic computations of EFGs. Benchmark density functional calculations are reported for a large set of main group diatomic molecules, a set of Cu and Au diatomics, several Ru and Nb complexes, the free uranyl ion, and two uranyl carbonate complexes. Data obtained from nonhybrid as well as fixed and range-separated hybrid functionals are compared. To allow for a chemically intuitive interpretation of the results, a breakdown of the EFGs of selected systems in terms of localized molecular orbitals is given. For CuF, CuCl, AuCl, UO₂²⁺, and a uranyl carbonate complex, the localized orbital decomposition demonstrates in particular the role of the valence metal d and f shells, respectively, and leads to rather compact analyses. For f orbitals, a Townes–Dailey-like model is set up to assist the analysis.

1. Introduction

In a molecule or in a crystal, a nuclear quadrupole moment interacts with the gradient of the electric field caused by the surrounding nuclei and electrons. This interaction is of high importance in experimental methods such as NMR,^{1,2} NQR (nuclear quadrupole resonance),^{1,3,4} Mössbauer spectroscopy,^{5,6} and other high-resolution spectroscopic techniques. Quadrupolar coupling interactions are routinely observed in solution NMR spectra in the form of line broadening arising from relaxation mechanisms.⁷ In solid-state NMR spectra, these interactions are found to dramatically alter spectral appearance. An extensive amount of work has been dedicated to the accurate calculation of nuclear quadrupole coupling parameters because of the importance of understanding their

relationship with molecular structure.^{8–11,2} From a quantum chemistry point of view, this structure–property relationship entails the accurate calculation of the electric field gradient (EFG).

The EFG is a molecular property that requires an all-electron treatment, at least for the atom of interest. When considering heavy atomic compounds, it is necessary to include relativistic effects in the calculations.^{2,11} One way to include those effects is through four-component relativistic methods.¹² However, the usage of those methods tends to be limited to the study of small systems because they are computationally expensive. An efficient method for computing EFGs in “relativistic” molecular systems (e.g., systems containing heavy atoms) is a quasi-relativistic four-component density reconstructed electric field gradient method. Such a method was introduced by van Lenthe and Baerends¹³ in the framework of the two-component zeroth-order regular approximation (ZORA) using density functional theory (DFT) and a Slater-type atomic orbital (STO) basis set. To

* To whom correspondence should be addressed. E-mail: jochena@buffalo.edu.

[†] State University of New York at Buffalo.

[‡] Pacific Northwest National Laboratory.

eliminate the bulk of picture-change effects, a reconstruction of the four-component electron density was proposed both in a scalar relativistic approximation (SRZ4) and in a variant including spin–orbit coupling (Z4). The approach was shown to provide reasonable results for EFG in a variety of molecules at an affordable computational expense. More recently, a similar approach has also been applied with ZORA and the two-component Douglas–Kroll–Hess (DKH) method using Gaussian type orbital (GTO) basis functions,¹⁴ as well as explicit perturbation expansions of the DKH operator,^{14,15} albeit only in a scalar relativistic framework. Among the available perturbative relativistic approaches, we mention an application of a scalar relativistic flavor of direct relativistic perturbation theory (DPT)^{16,17} to calculations of EFGs in bromofluoromethane.¹⁸

The present work reports an implementation for EFG calculations within the Z4 and SR-Z4 frameworks, utilizing the recently developed ZORA module of the NWChem package,¹⁹ which employs GTO basis sets for molecular calculations. The availability of this new implementation allows for the investigation of some questions regarding EFG calculations for main group elements and transition metals, among those the performance of the Z4 calculations in conjunction with range-separated hybrid functionals^{20–22} in DFT that have recently been implemented in NWChem,^{23,24} and the origin of the extreme sensitivity of copper and gold EFGs to approximations in the density functional.²⁵ We have developed a localized molecular orbital (LMO) analysis tool for field gradients to investigate the sensitivity of Cu and Au EFGs. The LMO analysis is also applied to study the effect of equatorial coordination on the uranium EFG in uranyl. Another aim of this paper is to test the suitability of various GTO basis set combinations for routine computations of heavy atom EFGs. Many of the previous works cited above used customized basis sets that are not readily available, for example, at the EMSL basis set exchange. We demonstrate by comparison with accurate literature data that suitable results—within typical error bars of DFT calculations—can be obtained for heavy atom EFGs with the basis sets used in this work, and that the LMO decomposition of EFGs leads to quite compact and chemically intuitive interpretations.

The paper is organized as follows: In section 2, the Z4 formalism for EFGs used for the implementation is outlined. Section 3 reports computational details and a few details about the program implementation. Results are presented and discussed in section 4 in the following order: main group diatomics, Cu and Au diatomics, complexes with Nb and Ru, and uranyl salts. Section 5 summarizes the findings.

2. Methodology

The electric field gradient (EFG) is evaluated from $\nabla\mathbf{E}(\mathbf{r})$ where the electric field is $\mathbf{E}(\mathbf{r}) = -\nabla\Phi(\mathbf{r})$. Here, $\Phi(\mathbf{r})$ is the electrostatic potential in an atom or molecule at point \mathbf{r} caused by the distribution of charges (electrons, nuclei, and the surrounding of the system). Thus,

$$\text{EFG}^{pq}(\mathbf{r}) = -\frac{\partial^2\Phi(\mathbf{r})}{\partial r_p\partial r_q} = -V_{pq}(\mathbf{r}) \quad (1)$$

where $r_p, r_q \in \{x, y, z\}$. The tensor \mathbf{V} in its principal axis system, $V_{ij}(\mathbf{r})$ ($i, j \in \{1, 2, 3\}$), is a definition of the EFG often used by solid state NMR spectroscopists. The nuclear quadrupole coupling constant C_Q , for instance, is determined by V_{33} , which is the principal component with the largest magnitude. The conversion between C_Q in MHz and the EFG is $C_Q = 234.9647QV_{33}$, with the nuclear quadrupole moment Q here given in barn (1 barn = 10^{-28} m²) units and V_{33} in atomic units (a.u.). The electric potential $\Phi(\mathbf{r})$ has nuclear and electronic contributions:

$$\Phi(\mathbf{r}) = \Phi_{\text{nuc}}(\mathbf{r}) + \Phi_{\text{el}}(\mathbf{r}) \quad (2)$$

with $\Phi_{\text{nuc}}(\mathbf{r})$ and $\Phi_{\text{el}}(\mathbf{r})$ defined as

$$\Phi_{\text{nuc}}(\mathbf{r}) = \sum_A \frac{Z_A}{|\mathbf{r} - \mathbf{R}_A|} \quad (3)$$

$$\Phi_{\text{el}}(\mathbf{r}) = -\int d^3r' \frac{\rho(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} \quad (4)$$

in atomic units (au). Here, Z_A is a nuclear charge, \mathbf{R}_A is a nuclear center, and $\rho(\mathbf{r})$ is the electronic charge density.

Accordingly, the EFG has nuclear and electron contributions, where the nuclear term, $V_{pq}^{\text{nuc}}(\mathbf{r})$, for point nuclei is

$$\begin{aligned} V_{pq}^{\text{nuc}}(\mathbf{r}) &= \frac{\partial^2}{\partial r_p\partial r_q} \left(\sum_A \frac{Z_A}{|\mathbf{r} - \mathbf{R}_A|} \right) \\ &= \sum_A Z_A \hat{Q}_{pq}(\mathbf{r}, \mathbf{R}_A) \end{aligned} \quad (5)$$

and the electronic term, $V_{pq}^{\text{el}}(\mathbf{r})$, is

$$\begin{aligned} V_{pq}^{\text{el}}(\mathbf{r}) &= \frac{\partial^2}{\partial r_p\partial r_q} \left(-\int d^3r' \frac{\rho(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} \right) \\ &= -\int d^3r' \rho(\mathbf{r}') \hat{Q}_{pq}(\mathbf{r}, \mathbf{r}') \end{aligned} \quad (6)$$

Here, the point quadrupole operator $\hat{Q}_{pq}(\mathbf{r}, \mathbf{r}')$ is given as

$$\hat{Q}_{pq}(\mathbf{r}, \mathbf{r}') = \frac{3(r_p - r'_p)(r_q - r'_q) - \delta_{pq}(\mathbf{r} - \mathbf{r}')^2}{|\mathbf{r} - \mathbf{r}'|^5} \quad (7)$$

The expression for the electronic component of V_{pq} may also be derived within a perturbation theory framework, as a derivative of the electronic energy of the system with respect to the perturbation by a small point quadrupole Q located at position \mathbf{r} , with the derivative taken at $Q = 0$. The perturbation operator is $-\hat{Q}_{pq}$. According to the Hellmann–Feynman theorem, the first-order perturbed energy is then $V_{pq}^{\text{el}} = -\langle\Psi|\hat{Q}_{pq}|\Psi\rangle$ in wave function theories, leading to expression 6.

In Kohn–Sham DFT, where the electron density is parametrized as $\rho = \sum_i \varphi_i^\dagger \varphi_i$, the energy functional minimized in the Kohn–Sham procedure may be written as

$$\tilde{E} = \sum_i \varepsilon_i - E_C - \int d^3r \cdot (\rho V_{XC}) + E_{XC} - \sum_i \varepsilon_i \left(\sum_{\mu\nu} C_{\mu i}^* S_{\mu\nu} C_{\nu i} - 1 \right) \quad (8)$$

where E_C and E_{XC} are the Coulomb and exchange-correlation (XC) energy functionals, and V_{XC} is the XC potential. We assume canonical MOs where the Lagrange multiplier ε_i is an eigenvalue of the Fock operator $\hat{F} = \hat{h} + V_C + V_{XC}$; that is, $\varepsilon_i = F_{ii} = \langle \varphi_i | \hat{F} | \varphi_i \rangle$. The last term with $S_{\mu\nu} = \langle \chi_\mu | \chi_\nu \rangle$ and the MO coefficients $C_{\mu i}$ ensures orthonormalization of the Kohn–Sham orbitals φ_i . Here and elsewhere, we assume single occupation of each molecular orbital (MO). Summation over spin indices is implied. A first order derivative of this functional with respect to a perturbation Q is

$$\frac{d\tilde{E}}{dQ} = \frac{\partial \tilde{E}}{\partial Q} + \sum_{\mu i} \frac{\partial \tilde{E}}{\partial C_{\mu i}} \cdot \frac{dC_{\mu i}}{dQ} = \frac{\partial \tilde{E}}{\partial Q} \quad (9)$$

since the energy is minimized with respect to variations of the MO coefficients. Thus, from eqs 8 and 9 for a basis set that is not dependent on the perturbation, one obtains

$$\frac{d\tilde{E}}{dQ} = \sum_i \frac{\partial \varepsilon_i}{\partial Q} = \sum_i \frac{\partial h_{ii}}{\partial Q} \quad (10)$$

With $-\hat{Q}_{pq}$ substituted for the operator derivative, the right-hand side yields eq 6. The definition of the EFG via a sum of orbital energy derivatives is used below when the scaled ZORA orbital energies are defined.

The equation for the electronic term of the EFG implies that ρ is the correct electronic charge density. This is the case in nonrelativistic theory, and in four-component relativistic methods. In the present work, we base the EFG calculations on the quasi-relativistic two-component ZORA method where the Kohn–Sham equations read

$$\left[(\boldsymbol{\sigma} \cdot \hat{\mathbf{p}}) \frac{\hat{K}}{2} (\boldsymbol{\sigma} \cdot \hat{\mathbf{p}}) + V(\mathbf{r}) \right] \varphi_i = \varepsilon_i^{\text{ZORA}} \varphi_i \quad (11)$$

with

$$\hat{K} = \frac{2c^2}{2c^2 - V} \quad (12)$$

The vector spin operator is $\boldsymbol{\sigma} = \sigma_x \hat{x} + \sigma_y \hat{y} + \sigma_z \hat{z}$. The σ_k ($k = x, y, z$) are the Pauli matrices, the linear momentum is given by $\hat{\mathbf{p}} = -i\nabla$, and the two-component molecular spinors $\varphi_i(\mathbf{r})$ are

$$\begin{aligned} \varphi_i(\mathbf{r}) &= \sum_{\mu} \chi_{\mu}(\mathbf{r}) \begin{pmatrix} C_{\mu i}^{\alpha} \\ C_{\mu i}^{\beta} \end{pmatrix} \\ &= \begin{pmatrix} \varphi_i^{\alpha}(\mathbf{r}) \\ \varphi_i^{\beta}(\mathbf{r}) \end{pmatrix} \end{aligned} \quad (13)$$

In the implementation used for the present work, the $\chi_{\mu}(\mathbf{r})$ are Gaussian-type AO basis functions (GTOs). The potential V includes the nuclear electron, the Coulomb, and the XC potential. In our implementation, V used in eq 12 is based on an electron density calculated at the Hartree–Fock (HF) level¹⁹ and excludes XC terms (unlike some other ZORA

implementations^{26–28}). This approximation has been carefully benchmarked in ref 19 and will be validated further in this work by comparisons of calculated EFGs with literature data.

The two-component electron density $\rho^Z = \sum_i \varphi_i^{\dagger} \varphi_i$ is not the charge density,²⁹ leading to picture-change errors in the calculated EFG.^{13–15,30,31} Van Lenthe and Baerends suggested reconstructing a normalized four-component electron charge density from the ZORA orbitals for the purpose of EFG computations in order to eliminate most of the picture-change errors in order c^{-2} . The electronic part of the EFG can be obtained from a four-component (Dirac) electron density, $\rho^D(\mathbf{r}) = \sum_i (U^{\dagger} \varphi_i(\mathbf{r}))^{\dagger} (U^{\dagger} \varphi_i(\mathbf{r}))$, where U is the Foldy–Wouthuysen unitary transformation.^{13,32} From using an approximation of U consistent with ZORA, the electron charge density is reconstructed, $\rho^D(\mathbf{r}) \approx \rho^{Z4}(\mathbf{r})$. The quasi-relativistic corrected normalized electron density, $\rho^{Z4}(\mathbf{r})$ in the context of ZORA is given by¹³

$$\rho^{Z4}(\mathbf{r}) = \sum_{i=1}^{\text{occ}} \frac{\rho_i^Z(\mathbf{r}) + \rho_i^S(\mathbf{r})}{1 + S_i} \quad (14)$$

where

$$S_i = \langle \varphi_i | (\boldsymbol{\sigma} \cdot \hat{\mathbf{p}}) \hat{B} (\boldsymbol{\sigma} \cdot \hat{\mathbf{p}}) | \varphi_i \rangle \quad (15)$$

and

$$\hat{B} = \frac{c^2}{[2c^2 - V]^2} \quad (16)$$

Here, the ZORA electron density, $\rho_i^Z(\mathbf{r})$, and the “small component” density, $\rho_i^S(\mathbf{r})$, for each MO are

$$\rho_i^Z(\mathbf{r}) = \varphi_i^{\dagger}(\mathbf{r}) \varphi_i(\mathbf{r}) \quad (17)$$

$$\rho_i^S(\mathbf{r}) = (\boldsymbol{\sigma} \cdot \hat{\mathbf{p}} \varphi_i)^{\dagger} \hat{B} (\boldsymbol{\sigma} \cdot \hat{\mathbf{p}}) \varphi_i \quad (18)$$

It is beneficial to adopt the “scaled ZORA” approach.^{33,34} The ZORA orbitals are usually excellent approximations of fully relativistic orbitals for valence and outer-core shells. However, orbital energies and properties calculated from the ZORA orbitals are very different from the fully relativistic results for core orbitals in heavy atoms. Van Lenthe and Baerends have shown that the orbital energies can be significantly improved toward the fully relativistic values upon applying of a scaling factor

$$\varepsilon_i^{\text{scaled-ZORA}} = \frac{\varepsilon_i^{\text{ZORA}}}{1 + S_i} \quad (19)$$

where S_i is the same term as in eq 15. Application of eq 10 with the *scaled* ZORA energies to derive the electronic EFG contribution leads to an EFG expression as in eq 6 with ρ replaced by the Z4 density (eq 14). Thus, the electronic term, $V_{pq}^{\text{el}}(\mathbf{r})$, reads in the Z4 method

$$\begin{aligned} V_{pq}^{\text{el}}(\mathbf{r}) &= - \int d^3r' \rho^{Z4}(\mathbf{r}') \hat{Q}_{pq}(\mathbf{r}, \mathbf{r}') \\ &= -I_{pq}^Z(\mathbf{r}) - I_{pq}^S(\mathbf{r}) \end{aligned} \quad (20)$$

In eq 20, $I_{pq}^Z(\mathbf{r})$ is given by

$$I_{pq}^Z(\mathbf{r}) = \sum_{\mu\nu} P'_{\mu\nu} \int d^3r' \chi_{\mu}(\mathbf{r}') \chi_{\nu}(\mathbf{r}') \hat{Q}_{pq}(\mathbf{r}, \mathbf{r}') \quad (21)$$

where $P'_{\mu\nu}$ represents a scaled density matrix:

$$P'_{\mu\nu} = \sum_{i=1}^{\text{occ}} \frac{C_{\mu i}^{\alpha*} C_{\nu i}^{\alpha} + C_{\mu i}^{\beta*} C_{\nu i}^{\beta}}{1 + S_i} \quad (22)$$

The MO integrals needed in eq 21 can be evaluated analytically in the AO basis in terms of Boys functions³⁵ or Rys polynomials³⁶ by standard procedures.^{37–41}

In eq 20, the term $I_{pq}^S(\mathbf{r})$ related to the small component density and the scaled-ZORA scaling factor can be calculated as

$$\begin{aligned} I_{pq}^S(\mathbf{r}) = & \sum_{\mu,\nu} P'_{\mu\nu} \sum_k \left\langle \frac{\partial \chi_{\mu}}{\partial r'_k} \middle| \hat{B} \hat{Q}_{pq}(\mathbf{r}, \mathbf{r}') \middle| \frac{\partial \chi_{\nu}}{\partial r'_k} \right\rangle + \\ & \sum_{\mu,\nu} P'_{x,\mu\nu} \left(\left\langle \frac{\partial \chi_{\mu}}{\partial y'} \middle| \hat{B} \hat{Q}_{pq}(\mathbf{r}, \mathbf{r}') \middle| \frac{\partial \chi_{\nu}}{\partial z'} \right\rangle - \right. \\ & \left. \left\langle \frac{\partial \chi_{\mu}}{\partial z'} \middle| \hat{B} \hat{Q}_{pq}(\mathbf{r}, \mathbf{r}') \middle| \frac{\partial \chi_{\nu}}{\partial y'} \right\rangle \right) + \\ & \sum_{\mu,\nu} P'_{y,\mu\nu} \left(\left\langle \frac{\partial \chi_{\mu}}{\partial z'} \middle| \hat{B} \hat{Q}_{pq}(\mathbf{r}, \mathbf{r}') \middle| \frac{\partial \chi_{\nu}}{\partial x'} \right\rangle - \right. \\ & \left. \left\langle \frac{\partial \chi_{\mu}}{\partial x'} \middle| \hat{B} \hat{Q}_{pq}(\mathbf{r}, \mathbf{r}') \middle| \frac{\partial \chi_{\nu}}{\partial z'} \right\rangle \right) + \\ & \sum_{\mu,\nu} P'_{z,\mu\nu} \left(\left\langle \frac{\partial \chi_{\mu}}{\partial x'} \middle| \hat{B} \hat{Q}_{pq}(\mathbf{r}, \mathbf{r}') \middle| \frac{\partial \chi_{\nu}}{\partial y'} \right\rangle - \right. \\ & \left. \left\langle \frac{\partial \chi_{\mu}}{\partial y'} \middle| \hat{B} \hat{Q}_{pq}(\mathbf{r}, \mathbf{r}') \middle| \frac{\partial \chi_{\nu}}{\partial x'} \right\rangle \right) \end{aligned} \quad (23)$$

where partial integration has been applied to avoid derivatives of \hat{B} . In the last equation, $P'_{x,\mu\nu}$, $P'_{y,\mu\nu}$, and $P'_{z,\mu\nu}$ are scaled spin-density matrices given by

$$P'_{x,\mu\nu} = i \sum_{i=1}^{\text{occ}} \frac{C_{\mu i}^{\alpha*} C_{\nu i}^{\beta} + C_{\mu i}^{\beta*} C_{\nu i}^{\alpha}}{1 + S_i} \quad (24)$$

$$P'_{y,\mu\nu} = \sum_{i=1}^{\text{occ}} \frac{C_{\mu i}^{\alpha*} C_{\nu i}^{\beta} - C_{\mu i}^{\beta*} C_{\nu i}^{\alpha}}{1 + S_i} \quad (25)$$

$$P'_{z,\mu\nu} = i \sum_{i=1}^{\text{occ}} \frac{C_{\mu i}^{\alpha*} C_{\nu i}^{\alpha} - C_{\mu i}^{\beta*} C_{\nu i}^{\beta}}{1 + S_i} \quad (26)$$

Those matrices have the property

$$P'_{k,\nu\mu} = -P'_{k,\mu\nu} \quad k = x, y, z \quad (27)$$

which can be used to prove that $I_{pq}^S(\mathbf{r})$ is a real quantity, as it should be.

The EFG from scalar relativistic ZORA is obtained by neglecting spin-dependent terms in the ZORA Kohn–Sham operator and in the EFG related integrals. The scaling integral, S_i , is redefined in the following way:

$$\begin{aligned} S_i &= \langle \varphi_i | (\boldsymbol{\sigma} \cdot \hat{\mathbf{p}}) \hat{B}(\boldsymbol{\sigma} \cdot \hat{\mathbf{p}}) | \varphi_i \rangle \\ &= \langle \varphi_i | \hat{\mathbf{p}} \hat{B} \hat{\mathbf{p}} + i \boldsymbol{\sigma} \cdot (\hat{\mathbf{p}} \hat{B} \times \hat{\mathbf{p}}) | \varphi_i \rangle \quad (28) \\ S_{\text{SR},i}^{\text{ZORA}} &= \langle \varphi_i | \hat{\mathbf{p}} \hat{B} \hat{\mathbf{p}} | \varphi_i \rangle \end{aligned}$$

As a consequence of the redefinition of the small component density, $\rho_i^S(\mathbf{r})$, the expression for $I_{pq}^S(\mathbf{r})$ is approximated as

$$I_{pq}^S(\mathbf{r}) \approx \sum_{\mu,\nu} P'_{\mu\nu} \sum_k \left\langle \frac{\partial \chi_{\mu}}{\partial r'_k} \middle| \hat{B} \hat{Q}_{pq}(\mathbf{r}, \mathbf{r}') \middle| \frac{\partial \chi_{\nu}}{\partial r'_k} \right\rangle \quad (29)$$

with $P'_{\mu\nu}$ now calculated with the scalar relativistic S_i from eq 28. The scalar relativistic $I_{pq}^Z(\mathbf{r})$ is calculated from eq 21 using the scalar relativistic $P'_{\mu\nu}$ terms.

3. Computational and Implementation Details

The quasi-relativistic EFG has been added to the recently implemented ZORA code reported by Nichols et al.¹⁹ in the NWChem 5.1 package.^{42–44} NWChem previously had a spin-free module for the evaluation of electric field gradients which used as electronic density $\rho^Z(\mathbf{r}) = \sum_i^{\text{occ}} \rho_i^Z(\mathbf{r})$. For this work, we added picture-change corrections,¹³ by choosing $\rho^{\text{ZA}}(\mathbf{r})$, eq 14, and extended the EFG module for spin–orbit Z4 computations.

The implementation of quasi-relativistic corrections to the EFG involved (i) evaluation of the scaled density matrices $P'_{\mu\nu}$, $P'_{x,\mu\nu}$, $P'_{y,\mu\nu}$, and $P'_{z,\mu\nu}$; (ii) evaluation of eq 21, $I_{pq}^Z(\mathbf{r})$; (iii) numerical computation of the integrals in eq 23, which allows for the evaluation of $I_{pq}^S(\mathbf{r})$. The evaluation of those integrals depends on the use of a reliable molecular integration grid. The grid was checked for the set of all diatomics by calculating the integrals of eq 21 analytically and by numerical integration. The results agreed within 5×10^{-3} relative error for the electronic component of V_{33} in the worst case (AuBr ZORA-4 CAM-B3LYP), and within 9×10^{-5} relative error on average. The evaluation of the electronic component of the EFG, $V_{pq}^{\text{el}}(\mathbf{r})$, followed steps i, ii, and iii; from there, $V_{pq}(\mathbf{r})$ was computed. The new module allows for two types of calculations: (a) restricted and unrestricted scalar relativistic ZORA: the module determines the EFG for scalar relativistic ZORA along with Z4 corrections (SR-Z4), suited for studying closed and open shell molecular systems; (b) spin–orbit ZORA and the associated Z4 corrections.

For the evaluation of the corrected electric field gradients at the heavy atom centers of interest, a choice of basis set with the following characteristics was adopted. (a) We chose a relativistic atomic GTO basis set that has the flexibility to include relativistic effects in the evaluation of the EFG integrals and is able to describe relativistic effects on the MOs. (b) Valence polarization functions were added where needed. The polarization functions were taken from the (def2) TZVPP⁴⁵ basis set available at the EMSL basis set exchange.^{46,47} (c) For increased flexibility in the core, the basis sets for the EFG atom were uncontracted. The basis sets chosen in this way represent a relatively economic choice

for routine EFG computations, somewhat similar in flexibility to the uncontracted Slater-type orbital (STO) basis sets chosen by van Lenthe and Baerends for their ZORA EFG computations.^{13,48}

For the section that validates the implementation with calculations of EFGs in diatomic molecules, we uncontracted the basis set from Tsuchiya–Abe–Nakajima–Hirao⁴⁹ (TANH) and added polarization functions from TZVPP. This basis set was also used for the calculation of EFGs in the copper diatomics. For the calculation of EFGs in gold diatomics, we used the SARC-ZORA TZVPP basis⁵⁰ for the Au atom and the uncontracted version of the TANH basis set along with polarization functions from TZVPP for all other atoms. In the “applications” sections, the basis sets used for the metal complexes were as follows: (i) ANO-RCC⁵¹ was used for ruthenium and niobium along with the 6-31G* basis set for ligand atoms. (ii) For comparison, calculations were also performed with the uncontracted TANH basis set polarized with TZVPP for Nb, and some comparisons were made by varying the ligand basis set as detailed in that section. (iii) For the uranium systems, we used the ANO-RCC basis set excluding *h* functions (see the section on metal complexes for benchmarks), along with TZVPP and 6-31G* for the ligand atoms. Some benchmark calculations in section 4.1 have also been carried out with the ANO basis set for heavy main group atoms because of convergence problems with the uncontracted TANH basis for At.

For the validation tests of EFGs for diatomic halides as well as the Cu and Au diatomics, we investigated the performance of the following functionals: Becke88⁵²+Perdew86⁵³ (BP), Becke88⁵²+LYP⁵⁴ (BLYP), Becke three-parameter Lee–Yang–Parr (B3LYP),⁵⁵ and two parametrizations of the Coulomb-attenuated (range-separated) version of B3LYP:²¹ CAM-B3LYP-A is the original parametrization with $\alpha + \beta = 0.65$, while CAM-B3LYP-B uses $\alpha + \beta = 1$ and is therefore fully long-range-corrected. We have also tested another parametrization of CAM-B3LYP which was optimized for EFG calculations on copper, CAM-B3LYP*.⁵⁶ For the study of larger systems, namely, the ruthenium, niobium, and uranium complexes, we used CAM-B3LYP-A, B3LYP, and revPBE.^{57–59} The revPBE and B3LYP functionals were chosen to facilitate a comparison with literature data for the Nb and Ru EFGs, and the comparison with CAM-B3LYP allows for an assessment of the performance of a range-separated functional for EFGs in these larger metal complexes.

For the localized orbital analysis of the EFGs, the following sequence of SR-Z4 calculations was employed in scalar relativistic calculations:

- (i) The EFG tensor (electronic plus nuclear contributions) was calculated and the principal axis system (PAS) determined.
- (ii) AO matrix elements used for the calculation of I_{pq}^Z and I_{pq}^S in eq 20 were combined. Corresponding EFG AO matrix elements for the nuclear EFG contributions were calculated from the AO overlap matrix, and from the AO matrix elements used to calculate the scalar relativistic S_i of eq 15, both scaled with V_{pq}^{nuc}/N . The electronic and “nuclear” AO matrices were added to

yield an AO matrix $h_{\mu\nu}^{pq}$ which yields the full (electronic plus nuclear) EFG tensor in the laboratory frame upon contraction with $P'_{\mu\nu}$ of eq 22.

- (iii) The $h_{\mu\nu}^{pq}$'s were transformed to the PAS to yield $h_{\mu\nu}^{11}$, $h_{\mu\nu}^{22}$, and $h_{\mu\nu}^{33}$. For highly symmetric molecules, this is not necessary, but it considerably simplifies the analysis for systems where the PAS does not coincide with the laboratory coordinates.
- (iv) The NBO 5.0 code version 5.0⁶⁰ was used to create a set of scalar relativistic natural localized MOs (NLMOs) and natural bond orbitals (NBOs). The columns of the NLMO to MO transformation matrix were scaled with the S_i , and the result was used together with the MO coefficients and the AO to NBO transformation to partition the principal components of the EFG tensor into contributions from individual NLMOs or NBOs. We refer the reader to refs 2, 61, and 62 for further details about the NBO/NLMO analysis of EFG tensors and other molecular properties. The implementation was originally carried out within the Amsterdam Density Functional package. For the purpose of this work, the analysis code has been turned into a stand-alone program that reads as input the AO matrices $h_{\mu\nu}^{11}$, $h_{\mu\nu}^{22}$, and $h_{\mu\nu}^{33}$; the scaling factors S_i ; the MO coefficients and the AO overlap matrix; and a set of matrices generated by the NBO program (transformations between NLMOs and MOs and between AOs and NBOs).

4. Results and Discussion

In this section, the implementation of the corrected electric field gradients in NWChem is validated by comparing a set of calculations on diatomic halides as well as Cu and Au diatomics with literature data. As an application to larger systems, we compute metal Z4 EFGs of a set of ruthenium, niobium, and uranium complexes. The results for these systems are going to be compared with previous calculations performed at the nonhybrid DFT level.

4.1. Test Set 1: Diatomic Main Group Halides. Table 1 lists EFGs at halides using the BP functional for SR ZORA, SR-Z4, ZORA, and Z4. The BP functional was used because of the availability of literature data with this functional from two different sources. The EFGs are in good agreement with the corresponding values in parentheses taken from the work of Neese et al.¹⁴ for F and At and from the work of van Lenthe and Baerends¹³ for the other molecules. Minor differences must be expected mainly because of the different basis sets used, and because of the different numerical integration grids applied in the DFT calculations. On a more technical level, there are also differences in how the potential V is included in the ZORA operator \hat{K} in eq 12, but prior benchmarks^{19,63,64} have indicated that the approximations in V made to construct \hat{K} are not overly critical beyond inclusion of the nuclear potential and a per atom electronic Coulomb potential.

In all cases, the sign and magnitude of the Z4 correction terms calculated for this work are in agreement with the literature data. The sign and magnitude of the spin–orbit

Table 1. Calculated Electric Field Gradient, V_{33} , at F, Cl, Br, I, and At in Hydrogen Halides and TII^a

	SR ZORA		SR-Z4		ZORA		Z4	
HF	2.7654	(2.7826)	2.7638	(2.7806)	2.7654		2.7638	
HCl	3.5561	(3.5627)	3.5481	(3.5481)	3.5562	(3.5627)	3.5485	(3.5533)
HBr	7.5752	(7.6182)	7.4569	(7.4706)	7.5835	(7.6295)	7.4970	(7.5430)
HI ^b	11.5014	(11.5959)	11.0035	(11.0470)	11.5556	(11.6700)	11.2335	(11.3369)
HAt	25.1128	(25.4477)	22.1853	(22.5743)	25.4021		23.4414	
TII ^b	2.2887	(2.3118)	2.1903	(2.2032)	2.7366	(2.7855)	2.7000	(2.7491)

^a BP functional, ANO-RCC basis for heavier atoms, and TZVPP for hydrogen. EFGs in parentheses are from the work of Neese et al.¹⁴ for F and At and from van Lenthe and Baerends¹³ for Cl, Br, and I. ^b For comparison, the results with the TANH basis (see section 3 and Table 2) are, for HI, SR-Z4 = 11.0590 and Z4 = 11.2893, and for TII, SRZ4 = 2.1881 and Z4 = 2.6987.

Table 2. Calculated Electric Field Gradient, V_{33} , at the Halide Nucleus for a Set of Diatomics^a

		NR		SR-Z4		Z4		observed
AlCl	³⁵ Cl	0.4168	(0.4153)	0.4206	(0.4193)	0.4207	(0.4201)	0.4602 ^b
GaCl	³⁵ Cl	0.6576	(0.6536)	0.6667	(0.6630)	0.6684	(0.6662)	0.6880 ^c
InCl	³⁵ Cl	0.6831	(0.6807)	0.7043	(0.7021)	0.7160	(0.7151)	0.6933 ^c
CuCl	³⁵ Cl	2.2801	(2.0798)	2.4885	(2.2909)	2.4889	(2.2945)	1.675 ^b
HCl	³⁵ Cl	3.5427	(3.5038)	3.5865	(3.5481)	3.5870	(3.5533)	3.516 ^d
ICl	³⁵ Cl	4.5996	(4.5656)	4.6844	(4.6453)	4.5697	(4.5353)	4.472 ^d
BrCl	³⁵ Cl	5.3737	(5.3375)	5.4355	(5.3975)	5.4159	(5.3845)	5.336 ^d
AlBr	⁷⁹ Br	0.9444	(0.9695)	1.0113	(1.0324)	1.0154	(1.0463)	1.112 ^e
GaBr	⁷⁹ Br	1.3133	(1.3335)	1.3995	(1.4130)	1.4086	(1.4343)	1.50 ^c
InBr	⁷⁹ Br	1.3784	(1.3988)	1.4848	(1.4981)	1.5172	(1.5407)	1.57 ^c
CuBr	⁷⁹ Br	4.5031	(4.2631)	5.1641	(4.9029)	5.1731	(4.9454)	3.71 ^f
HBr	⁷⁹ Br	7.0599	(7.0394)	7.5154	(7.4706)	7.5424	(7.5430)	7.55 ^d
IBr	⁷⁹ Br	9.3799	(9.3957)	9.9862	(9.9575)	9.8467	(9.8767)	9.89 ^c
BrCl	⁷⁹ Br	11.6139	(11.6457)	12.2909	(12.2841)	12.2761	(12.3437)	12.4 ^d
All	¹²⁷ I	1.4295	(1.4532)	1.6899	(1.7011)	1.7289	(1.7511)	1.90 ^c
Gal	¹²⁷ I	1.7598	(1.7869)	2.0515	(2.0706)	2.1024	(2.1348)	2.28 ^c
InI	¹²⁷ I	1.8682	(1.8911)	2.1862	(2.1884)	2.2763	(2.2914)	2.38 ^c
TII	¹²⁷ I	1.8774	(1.8979)	2.1881	(2.2032)	2.6987	(2.7491)	2.70 ^c
CuI	¹²⁷ I	6.2237	(5.9213)	7.7206	(7.3770)	7.7909	(7.4942)	5.79 ^f
HI	¹²⁷ I	9.5711	(9.5913)	11.0590	(11.0470)	11.2893	(11.3369)	11.3 ^d
IBr	¹²⁷ I	14.7489	(14.8218)	16.8048	(16.8388)	16.7654	(16.8696)	16.8 ^d
ICl	¹²⁷ I	15.7324	(15.8272)	17.9269	(17.9799)	17.7673	(17.8874)	18.1 ^d
IF	¹²⁷ I	18.4959	(18.6275)	21.0265	(21.1071)	20.7241	(20.8850)	21.2 ^d
AlF	²⁷ Al	-1.1148	(-1.1032)	-1.1194	(-1.1075)	-1.1194	(-1.1078)	-1.096 ^c
AlCl	²⁷ Al	-0.9199	(-0.9087)	-0.9221	(-0.9107)	-0.9221	(-0.9110)	-0.8828 ^c
AlBr	²⁷ Al	-0.8546	(-0.8436)	-0.8511	(-0.8402)	-0.8504	(-0.8399)	-0.8130 ^e
All	²⁷ Al	-0.7970	(-0.7865)	-0.7852	(-0.7748)	-0.7814	(-0.7714)	-0.7417 ^c
GaF	⁶⁹ Ga	-2.6053	(-2.6237)	-2.6846	(-2.6942)	-2.6937	(-2.7112)	-2.76 ^c
GaCl	⁶⁹ Ga	-2.2775	(-2.2866)	-2.3462	(-2.3472)	-2.3546	(-2.3627)	-2.38 ^g
GaBr	⁶⁹ Ga	-2.1577	(-2.1705)	-2.2131	(-2.2177)	-2.2195	(-2.2312)	-2.24 ^c
Gal	⁶⁹ Ga	-2.0507	(-2.0599)	-2.0830	(-2.0846)	-2.0813	(-2.0901)	-2.10 ^c
InF	¹¹⁵ In	-3.7354	(-3.7418)	-4.0378	(-4.0437)	-4.1338	(-4.1438)	-4.18 ^c
InCl	¹¹⁵ In	-3.4111	(-3.4140)	-3.6866	(-3.6889)	-3.7787	(-3.7826)	-3.79 ^c
InBr	¹¹⁵ In	-3.2887	(-3.2966)	-3.5425	(-3.5468)	-3.6296	(-3.6365)	-3.65 ^c
InI	¹¹⁵ In	-3.1791	(-3.1851)	-3.3956	(-3.3984)	-3.4685	(-3.4749)	-3.49 ^c

^a BP functional and TANH basis for heavy atoms. The EFGs in parentheses and the observed EFGs are from the work of van Lenthe and Baerends.¹³ References to the sources of the experimental data are given as additional footnotes. ^b Ref 65. ^c Ref 66. ^d Ref 67. ^e Ref 68. ^f Ref 69. ^g Ref 70.

corrections to the EFG also agree well with the data from ref 13. TII exhibits a particularly large spin-orbit effect on the iodine EFG (a 23% increase relative to the scalar Z4 value calculated in this work compared to 25% as reported by van Lenthe and Baerends¹³). The Z4 correction terms are largest for the heaviest nucleus among the set of molecules in Table 1, astatine, which is not surprising given the form of the operator \hat{B} in the correction terms. The correction of -3 atomic units is in agreement with the one predicted by Neese et al.¹⁴ As shown by our spin-orbit computations, the spin-orbit effect on the At EFG in HAt is small in comparison, yielding an increase of about 6%. The corresponding spin-orbit coupling induced increase of the iodine EFG in HI is only 2%.

Table 2 lists calculated EFGs at indicated atom centers for a larger set of main group diatomic halides obtained in nonrelativistic (NR), scalar relativistic Z4 (SR-Z4), and spin-orbit Z4 (Z4) computations. The BP functional was used again in order to facilitate a comparison with the literature data (in parentheses¹³). In this table, we also include experimentally observed EFGs for comparison. The values were taken from the work of van Lenthe and Baerends¹³ and originally published in the form of measured nuclear quadrupole coupling constants in refs 65-69. The number of significant figures provided for the experimental EFG relates to how accurately the nuclear quadrupole moments are known.⁷¹

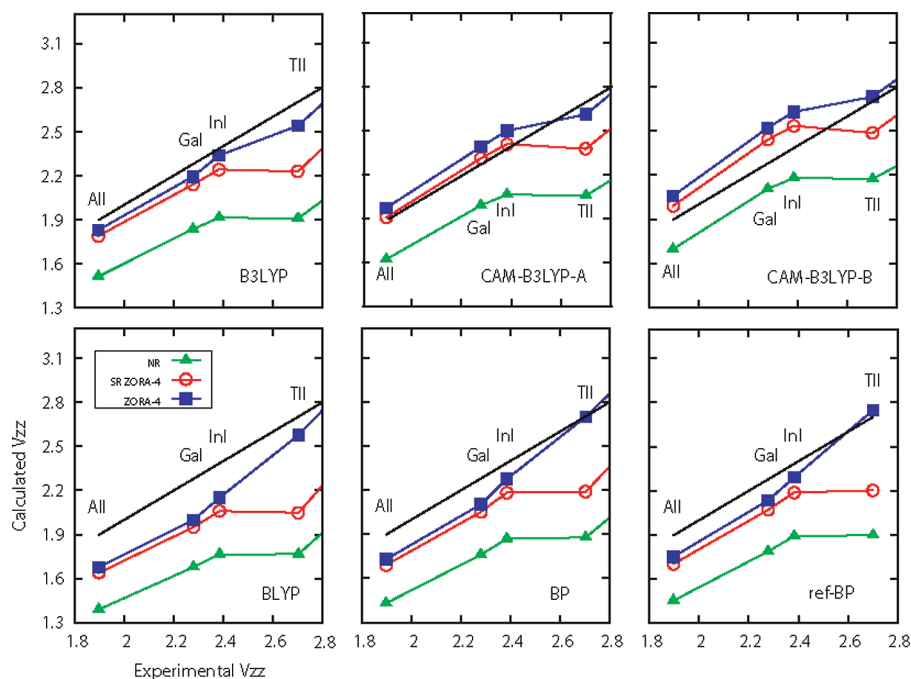


Figure 1. Calculated NR, SR-Z4, and Z4 iodine V_{33} vs experimental values, for different functionals. Atomic units. Straight black lines indicate where calc. = expt. The graph labeled ref-BP shows data from van Lenthe and Baerends¹³ for comparison.

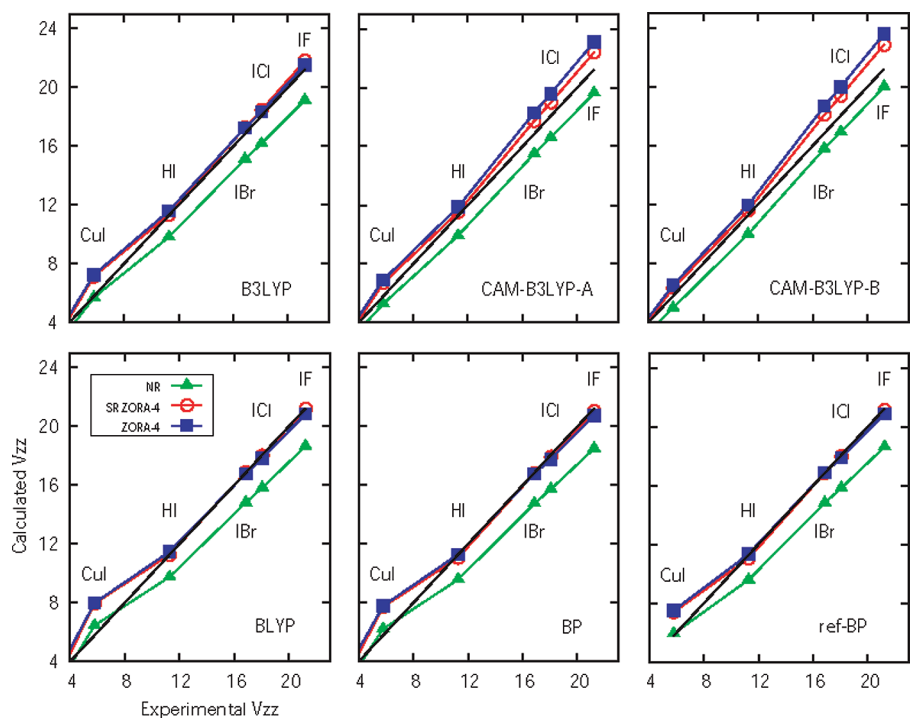


Figure 2. Calculated NR, SR-Z4, and Z4 iodine V_{33} vs experimental values, for different functionals. See also caption of Figure 1.

It is important to keep in mind some of the technical differences between the EFG calculations performed for this work and those from ref 13: van Lenthe and Baerends used Slater-type basis sets, while we used Gaussian-type basis sets. The numerical integration grids are quite different, and the implementations of the ZORA operator differ as well. It is therefore reassuring that Table 2 demonstrates close agreement between our results and those from ref 13 despite the

technical differences between the calculations. In Figures 1–4, plots of calculated vs experimental EFGs are shown. We have included calculated data from van Lenthe and Baerends for comparison to demonstrate the close agreement between the different calculations. The agreement with experimental results is generally good, with the exception of the Cu diatomics for which the agreement of the halide EFGs with experimental values is fair. This problem has

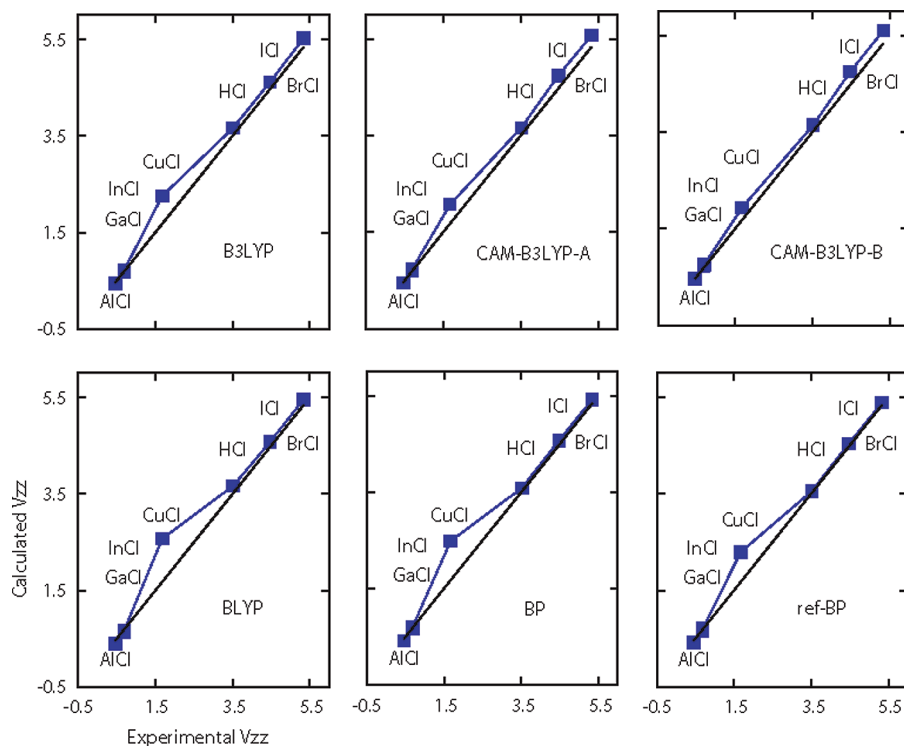


Figure 3. Calculated NR, SR-Z4, and Z4 chlorine V_{33} vs experimental values, for different functionals. See also caption of Figure 1.

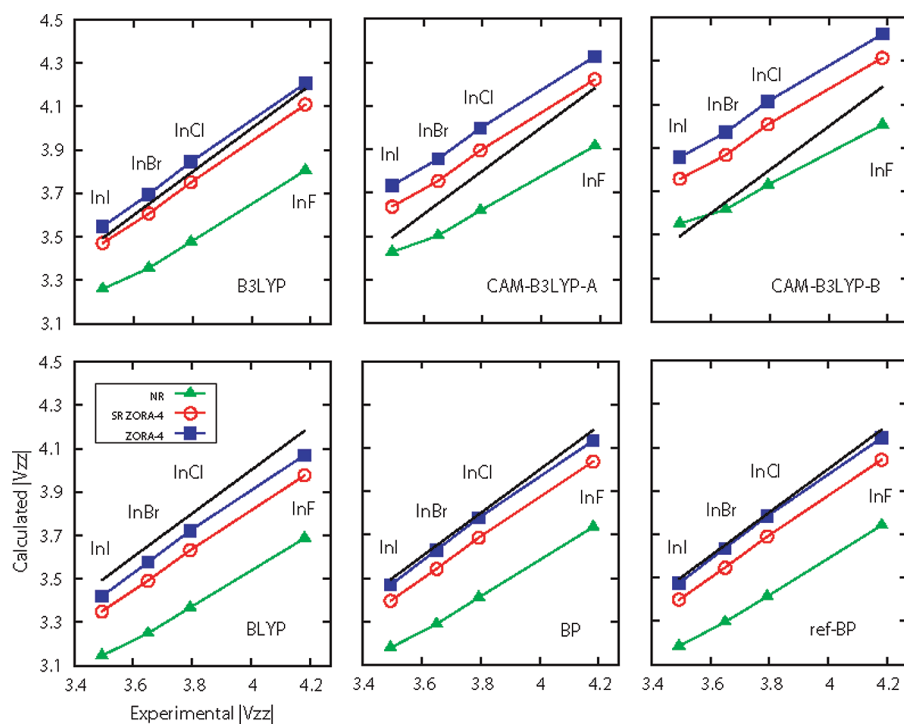


Figure 4. Calculated NR, SR-Z4, and Z4 indium V_{33} (absolute value) vs experimental values, for different functionals. See also caption of Figure 1.

already been noted in ref 13. In these molecules, the Cu EFG calculated with standard functionals even has the wrong sign, a problem that will be discussed in more detail below.

Figures 1–4 also show the dependence of the calculated EFGs when comparing different density functionals. The corresponding numerical data are provided in Tables 11–15 in the Appendix. In addition to BP, the functionals BLYP,

B3LYP, and two different parametrizations of CAM-B3LYP are compared for different levels of theory (NR, SR-Z4, and Z4). Figure 1 shows EFGs at iodine for AlI, GaI, InI, and TlI. An important feature of this graph is the behavior of EFG at iodine in TlI, where we observe closer agreement to the experimental value in the following order: NR, SR-Z4, and Z4. This trend holds for all of the functionals studied,

Table 3. Copper and Gold V_{33} , in Atomic Units, Calculated for Selected Diatomics with Z4 and Different Functionals^a

method ^b	CuH		CuF		CuCl		CuBr		CuI	
BLYP	0.800	(0.776)	0.916	(0.879)	0.731	(0.718)	0.699	(0.696)	0.633	(0.640)
B3LYP	0.486	(0.338)	0.372	(0.112)	0.307	(0.149)	0.312	(0.181)	0.313	(0.208)
CAMB3LYP	0.345	(0.311)	0.190	(0.145)	0.143	(0.129)	0.158	(0.152)	0.173	(0.176)
CAMB3LYP*	-0.048	(-0.085)	-0.461	(-0.503)	-0.303	(-0.315)	-0.231	(-0.237)	-0.146	(-0.141)

method ^c	AuF		AuCl		AuBr		AuI	
BLYP	4.018	(4.005)	3.592	(3.659)	3.485	(3.625)	3.250	(3.445)
B3LYP	2.290	(2.190)	2.208	(2.206)	2.213	(2.280)	2.220	(2.310)
CAMB3LYP	1.571	(1.285)	1.562	(1.384)	1.569	(1.502)	1.639	(1.625)
CAMB3LYP*	-0.512	(-0.858)	-0.058	(-0.290)	0.113	(-0.010)	0.358	(0.332)

^a The values in parentheses are four-component DFT results from Thierfelder et al.⁵⁶ ^b TANH basis set⁴⁹ for all atoms except hydrogen where the TZVPP basis set⁴⁶ was used. Polarization functions were taken from TZVPP. ^c SARC-ZORA basis set⁵⁰ for Au and TANH basis set elsewhere; TZVPP for hydrogen. Polarization functions were taken from TZVPP, except for Au.

and it demonstrates the potential importance of considering spin-orbit relativistic effects in computations of heavy atom EFGs, in particular for heavy p block elements. Relativistic effects are also significant for the indium EFGs (Figure 4) and for the remaining iodine EFGs (Figure 2). For the Cl EFGs of Figure 3, the differences between SR-Z4, Z4, and NR were not very pronounced, and therefore we decided to show only the spin-orbit Z4 data set in the graphs.

Overall, the results for the different sets of molecules show that Z4 is in reasonable agreement with experimental results for all functionals that were considered, unlike the NR calculations. In all cases, our results with the BP functional are very similar to those obtained by van Lenthe and Baerends.¹³ As with the data in Table 1, minor differences are most likely due to the different basis sets (GTO vs STO, and the overall number of functions) and numerical integration grids. Good performance of nonhybrid DFT and DFT with standard hybrid functionals for main group atom EFGs has been noted before.⁷² Figures 1–4 indicate a slightly better performance of the B3LYP hybrid and the two range-separated CAM-B3LYP parametrizations compared to the nonhybrid functionals BP and BLYP for iodine and chlorine EFGs, whereas for the indium EFGs the results are closest to experimental results with the BP and the B3LYP functionals. For the indium data set, the CAM-B3LYP-B parametrization, which is fully long-range corrected, leads to an overestimation of the EFG magnitudes such that the NR data end up being closer to experimental data—clearly, these are “better” results for the wrong reasons.

4.2. Test Set 2: Copper and Gold Diatomics. Table 3 lists EFG data for Cu and Au in a set of diatomic molecules. The data in parentheses were obtained by Thierfelder et al. with four-component relativistic DFT and very flexible GTO basis sets designed for EFG computations.^{56,73} As already mentioned in section 3, we employed the uncontracted TANH basis with polarization functions from TZVPP for Cu and SARC-ZORA for Au. Transition metal EFGs are more challenging to compute than EFGs for main group atoms⁷²—at least for diatomic molecules. As the table shows, the dependence of the functional is rather dramatic. The CAM-B3LYP* parametrization yields the desired negative signs for the Cu and Au field gradients in these molecules,⁵⁶ while standard nonhybrid and hybrid functionals predict the EFGs with the wrong sign. Given the extreme sensitivity of

Table 4. CuCl: NLMO Analysis of Copper V_{33} (au)^a

NLMO	B3LYP	CAM	CAM*
$\sigma(\text{Cu-Cl})$	-0.443	-0.477	-0.461
$\Sigma 2p$	0.058	0.050	0.018
$\Sigma 3p$	-0.172	-0.243	-0.461
d_δ (ea.)	8.749	8.750	8.730
d_π (ea.)	-4.308	-4.309	-4.305
d_σ	-7.765	-7.825	-8.005
$\Sigma 3d$	1.109	1.057	0.845
Σ analysis	0.552	0.387	-0.059
total calcd	0.295	0.133	-0.304

^a SRZ4 calculations. CAM = CAM-B3LYP-A, CAM* = CAM-B3LYP*.

the EFGs to the functional, the agreement of our results with the four-component DFT data of Thierfelder et al.⁵⁶ can be considered good. For example, the deviations of up to 0.3 au between our data and those of ref 56 for the gold EFGs are within 10% of the range of functional dependence for each diatomic. The EFG ranges from AuF to AuI for each functional are also in reasonable agreement with literature data.

The sensitivity of the metal EFGs begs the question, how exactly is V_{33} resulting from a balance between contributions from different orbitals? To address this question, a NLMO analysis as detailed in section 3 has been carried out for the CuCl and CuF diatomics. A careful analysis by Schwerdtfeger et al.⁷⁴ of the electron density distribution of CuCl calculated at various levels of theory has previously demonstrated that this system affords a sensitive coupling of the valence and core charge densities which strongly affects calculated properties (EFG, but also the dipole moment). The NLMO data are provided in Tables 4 and 5, with contributions grouped by orbital type. To simplify the interpretation, the analyses were carried out at the SRZ4 level and, for the Cu EFGs, afford results similar to those of spin-orbit Z4. The NBO analysis indicated very little delocalization in both systems; therefore, a separate decomposition of the EFG into contributions from NBOs instead of NLMOs does not offer much additional insight, and we forego a separate discussion of the NBO contributions. In order to assist the discussion, we recall a few details about the role of atomic orbitals for EFGs.^{2,75} For a completely ionic bond (Cu^+Cl^-), the EFG vanishes as long as the AOs are not polarized and as long as they completely shield the nuclear charges from the bonding partner. For the EFG component along the bond,

Table 5. CuF: NLMO Analysis of Copper V_{33} (au)^a

NLMO	B3LYP	CAM	CAM*
$\sigma(\text{Cu}-\text{F})$	-0.485	-0.508	-0.493
$\Sigma 2p$	0.028	0.022	-0.023
$\Sigma 3p$	-0.314	-0.388	-0.699
d_δ (ea.)	8.771	8.773	8.738
d_π (ea.)	-4.357	-4.354	-4.342
d_σ	-7.457	-7.547	-7.812
$\Sigma 3d$	1.371	1.291	0.980
Σ analysis	0.600	0.417	-0.235
total calcd	0.351	0.173	-0.467

^a SRZ4 calculations. CAM = CAM-B3LYP-A, CAM* = CAM-B3LYP*.

V_{33} , with pure atomic orbitals, the d_σ orbital on Cu would contribute a relative increment of -2 to the EFG; the d_π orbitals, -1 each; and the d_δ , increments of $+2$ each.² For a filled atomic d shell, the contributions cancel. For an atomic p shell, the relative magnitudes are -2 for p_σ versus $+1$ for each p_π . The analysis for CuCl in Table 4 demonstrates that the Cu 3d shell and the 3p semicore shell both contribute significantly to the EFG and that the EFG caused by the Cu 3p and 3d shells accounts for most of the trend going from B3LYP to CAM-B3LYP*.

The d_δ orbitals are nonbonding. Their -8.75 to -8.73 au EFG contribution can therefore be taken as those of pure Cu 3d orbitals. It is reassuring that their EFG contribution remains almost constant between the calculations with the different functionals. For an essentially spherical Cu^+ d^{10} ion, one would then expect approximately 4.37 au from each d_δ and around -8.75 to -8.73 au from the d_σ orbital. The d_π EFG contributions are very close to the idealized value. The minor deviation can be attributed to π interactions between Cu and Cl. However, both for B3LYP and CAM-B3LYP-A, the negative contribution from d_σ is significantly lower than what would be expected for a pure Cu 3d, indicating that with these functionals the d_σ orbital is significantly more involved in the covalent Cu-Cl bond than in the CAM-B3LYP* calculation. The EFG trend for the d_σ reflects the NLMO compositions: For B3LYP, $\sigma(\text{Cu}-\text{Cl})$ has an overall 17% Cu character (the small value reflecting the ionicity of the bond) of which 5.7% is d_σ . For comparison, with CAM-B3LYP*, the bond is somewhat more ionic (12% Cu) but with only a 3.5% admixture of d_σ . The trends for the Cu 4s mixing into the d_σ NLMO go in the same direction: with B3LYP, there is 5.8% Cu 4s in the d_σ NLMO, whereas in the CAM-B3LYP* calculation, the 4s percentage is 3.4%. The effect on the EFG from $\sigma(\text{Cu}-\text{Cl})$ is only minor, but because of the strong EFG caused by even small changes in the occupations and spatial extent of the individual 3d orbitals, the effect on the EFG contribution from d_σ is large. The EFG contributions from the Cu 3p shell can be attributed to electrostatic polarization by the nonspherical environment and valence-core orthogonalization.

The CuF data in Table 5 show that this case is very similar to CuCl, except the bond is more ionic, as indicated by a smaller percentage of Cu orbitals in the $\sigma(\text{Cu}-\text{F})$ NLMOs (ranging from 10.9% with B3LYP to 6.8% with CAM-B3LYP*) compared to CuCl. We should point out that a breakdown of the EFG in CuF in terms of orbital contributions has also been given in ref 25, where the authors stated

Table 6. AuCl: NLMO Analysis of Gold V_{33} (au)^a

NLMO	B3LYP	CAM	CAM*
$\sigma(\text{Au}-\text{Cl})$	-1.577	-1.702	-1.696
Au core	0.451	0.388	0.136
f_σ	-30.512	-30.510	-30.556
f_π (ea.)	-22.884	-22.882	-22.915
f_δ (ea.)	0.010	0.010	0.010
f_ϕ (ea.)	38.185	38.181	38.234
$\Sigma 4f$	0.110	0.108	0.102
5s	-0.028	-0.030	-0.027
5p $_\sigma$	-249.412	-252.528	-250.934
5p $_\pi$ (ea.)	124.512	125.849	124.524
$\Sigma 5p$	-0.388	-0.830	-1.886
5d $_\sigma$	-11.370	-11.601	-11.883
5d $_\pi$ (ea.)	-7.513	-7.542	-7.571
5d $_\delta$ (ea.)	15.147	15.213	15.241
$\Sigma 5d$	3.898	3.741	3.457
Σ all other NLMOs	-0.135	-0.105	-0.131
Σ analysis	2.331	1.570	-0.045
total calcd	2.332	1.573	-0.040

^a SRZ4 calculations. CAM = CAM-B3LYP-A, CAM* = CAM-B3LYP*.

that the “total EFG for the Cu 3d shell adds up to 8.1, 13.3, and 12.3 au at the HF, LDA and B3LYP level of theory”, which was found to be balanced by contributions from fluorine 2p. The EFG and its functional dependence were mainly attributed to polarization of the Cu 3d shell. From our analyses on CuF and CuCl, we find that the contributions from the Cu 3d shell are definitely crucial, but it appears that hybridization, i.e., the extent of how much the Cu 3d $_\sigma$ participates in the bonding and how well the polarization of the 3p shell is described, are major factors deciding which functional performs best. It is known that for Cu(II) systems calculated with standard functionals the metal d shell binds too strongly covalently.⁷⁶ Although the Cu-X diatomics are very different systems, a similar problem might be manifesting here in the form of a too strong involvement of the Cu 3d $_\sigma$ orbital in the Cu-X bond.

For comparison, and to investigate relative contributions to the EFG from f orbitals, a NLMO decomposition was also carried out for the Au EFG in AuCl. The SRZ4 data are collected in Table 6. Regarding the sign and relative magnitude of EFG contributions from idealized atomic f shells (spin-free), see Figure 5. The f_δ orbitals have cubically symmetric charge densities and therefore do not contribute to the EFG at their atomic center. The calculated EFG contributions for the gold 4f shell listed in Table 6 nicely follow the ratios predicted from the simple AO model. For instance, the negative $4f_\sigma$ contribution ($V_{33} = V_{zz}$) is 1.333 times those from each $4f_\pi$ (idealized AOs: 4/3). The $4f_\delta$ contributions are effectively zero (the small EFG indicating a slight polarization of the 4f shell along with nonspherically symmetric valence-core polarization effects), and the ratio of the $4f_\phi$ to the $4f_\pi$ EFG terms of 1.669 for AuCl is also in excellent agreement with the idealized value of $-5/3$ from Figure 5. Overall, the 4f shell is not a dominant source of the EFG in AuCl when calculated with the B3LYP and CAM-B3LYP functionals. For CAM-B3LYP*, the calculated net EFG almost vanishes as a result of a cancellation of terms with opposite signs. The origin of the rather large EFGs calculated with B3LYP and CAM-B3LYP is seen to be similar to that for CuCl. A closer inspection of the NLMO

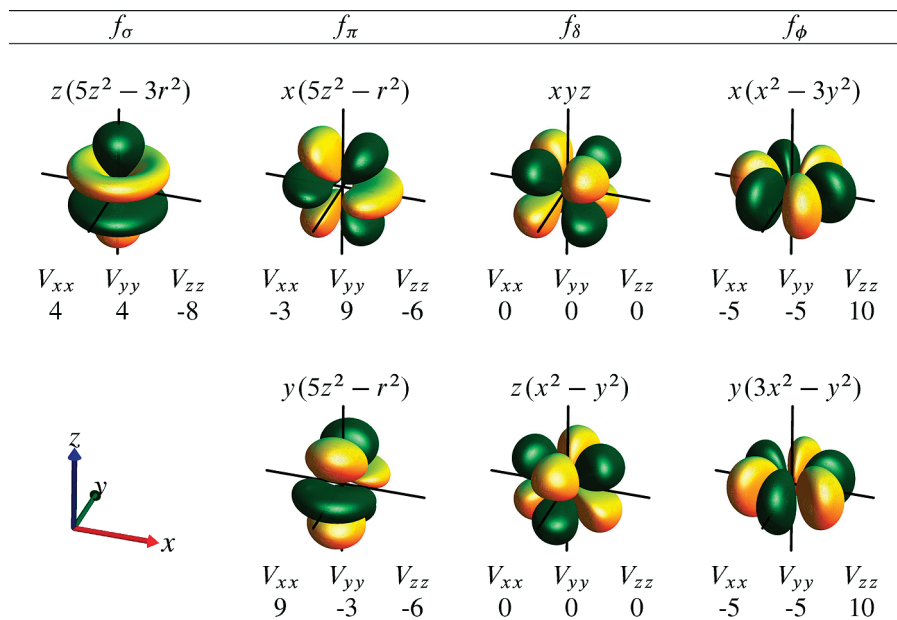


Figure 5. Sign and relative magnitude of EFG contributions from atomic f shells (spin-free), calculated analytically from 4f orbitals $N\Omega(x, y, z) e^{-\alpha r}$ and the nonrelativistic EFG operator, with Ω as indicated in the figures and N being a normalization constant. The EFG values listed are in units of $2\alpha^3/630$ per double occupation. Symmetry labels with respect to z axis.

compositions also showed that the trends going from B3LYP to CAM-B3LYP* are similar, with the bond becoming slightly more ionic, the d character in the bond NLMO decreasing, and the s character of the $5d_\sigma$ NLMO decreasing. The net positive EFG from 5d remains substantial, however. Core polarization (1s to 4d) and valence-core orthogonalization also play a role, as seen from the combined Au core terms in Table 6. As we have argued elsewhere,² in accord with other researchers,^{48,74} the core terms are coupled to the valence EFG terms. For AuCl, they follow in sign the overall trend seen for the functionals, i.e., the EFG becoming less positive when going from B3LYP to CAM-B3LYP*. The 5p semicore orbitals have very large EFG contributions individually, indicating that even a slight imbalance among these orbitals with respect to a spherically symmetric p shell can cause a sizable EFG. The combined 5p EFG contributions in AuCl follow the same trend as 5d and the core terms, becoming more negative when going from B3LYP to CAM-B3LYP*, and provide the largest individual fraction of the overall trend.

4.3. Ru and Nb Complexes. The present section focuses on metal EFGs in Ru and Nb complexes. Each of these complexes exhibits a particular, different role of the valence d shell population for the EFG, as analyzed recently.² There are also numerous important applications of complexes of this type, as well as fundamental interest from a structure and bonding viewpoint. For instance, the complex $[\text{RuCl}_4\text{N}]^-$ features a formal Ru–nitrogen triple bond and an unusual oxidation state. Ruthenocene, $\text{Ru}(\text{C}_5\text{H}_5)_2$, has long fascinated chemists. Among its applications, it has the ability to transfer electrons, which makes it a good photoinitiator for polymerization reactions.⁷⁷ There are numerous papers on $[\text{Ru}(\text{bipy}_3)]^{2+}$ and its derivatives, and similar complexes with bidentate unsaturated ligands, with proposed applications in fields as diverse as photovoltaics,^{78,79} materials science,^{80–83} biology,⁸⁴ biochemistry,⁸⁵ and medicine.^{85–89} The half-

Table 7. Computed Ru V_{33} , in Atomic Units, for $[\text{RuCl}_4\text{N}]^-$ Using Different Levels of Theory^a

	CAM-B3LYP-A	B3LYP	revPBE
ZORA/6-311G**	4.776	4.416	4.033
Z4/6-311G**	4.693	4.346	3.969
Z4/6-31G*	4.691	4.347	3.972
ZORA/ANO-(H set)	4.766	4.422	4.031
Z4/ANO-(H set)	4.683	4.350	3.968
ZORA/ANO-(G,H set)	4.806	4.455	4.052
Z4/ANO-(G,H set)	4.724	4.383	3.989

^a ANO-(G,H set) is ANO-RCC for Ru but with h or g , h functions removed from the basis. See text for details.

sandwich niobium metallocenes CpNbCl_4 and $\text{CpNb}(\text{CO})_4$ are interesting, for instance, because of their applications in catalysis.⁹⁰

We first investigate the basis set dependence of the metal EFGs. The ANO basis set used for the metal in these calculations contains functions with a high angular momentum that may serve as polarization functions but is also important in correlated wave function electronic structure methods to describe correlation. In DFT calculations, one might be able to exclude the h set and perhaps even the g set. There is also the question of how flexible the ligand basis set has to be. Table 7 provides calculated ruthenium EFGs in $[\text{RuCl}_4\text{N}]^-$ at the ZORA spin–orbit level, with and without Z4 corrections, for several functionals with the following basis sets: ANO for Ru and 6-31G* versus 6-311G** for the ligand atoms and ANO versus ANO-(H set) versus ANO-(G,H sets) along with 6-31G* to study the effect of removing g and h functions from the metal basis. From these calculations, we observe the following: (i) The effect of adding the Z4 corrections lowers the EFG values by about 2% for all the functionals tested. (ii) Using a more flexible basis set than 6-31G* for the ligands does not affect the EFG at Ru noticeably. This allows a significant reduction of the computational cost for larger systems. (iii) Removing

Table 8. Computed Spin–Orbit Z4 and Experimental Metal Atom V_{33} , in Atomic Units, for Selected Ru and Nb Complexes^a

	CAM-B3LYP-A	B3LYP	revPBE	lexptll ^b
[RuCl ₄ N] [−]	4.691	4.347 (4.33)	3.972 (3.94)	4.27(16)
Ru(C ₅ H ₅) ₂	1.917	1.659 (1.98)	1.377 (1.66)	1.15(5)
[Ru(bipy ₃)] ²⁺	−0.190	−0.283 (−0.23)	−0.378 (−0.36)	<0.21
CpNbCl ₄	1.406	1.312 (1.28)	1.278 (1.26)	0.72(5)
CpNbCl ₄ ^c	1.394	1.297	1.262	
CpNb(CO) ₄	0.225	0.240 (0.24)	0.274 (0.27)	0.01(0)
CpNb(CO) ₄ ^c	0.238	0.253	0.292	

^a The calculated EFGs in parentheses were taken from ref 2. ^b Experimental EFGs for ruthenium complexes were calculated in ref 91 from Mössbauer quadrupole splittings reported in ref 92. Experimental EFGs for the niobium complexes were calculated from the solid-state ⁹³Nb NMR quadrupole coupling constant of ref 93 using a nuclear quadrupole moment of $-0.320(20)$ barn.⁷¹ ^c Using uncontracted TANH basis set for niobium.

the *h* functions from the ANO basis for Ru has a negligible effect. When removing both the *g* and *h* sets, the EFG at Ru varies by less than 1% compared to the full ANO basis. Thus, it appears that in DFT computations of EFGs for the 4d metals, the high-angular momentum basis functions in the ANO basis are of little benefit. This is most likely due to the fact that electron correlation in DFT is obtained from the integration of a rather smooth electron density and its derivatives in the exchange–correlation functional, while in correlated wave function methods, high angular momentum basis functions are required for better approximating the electron cusps in the wave function.

Table 8 lists Z4 EFG data calculated for a set of ruthenium and niobium complexes using the CAM-B3LYP-A, B3LYP, and revPBE functionals. The calculated EFGs in parentheses were taken from ref 2 where SR-Z4 was used along with a Slater-type TZ2P basis set and optimized geometries. We employed the same optimized geometries for easier comparison. The technical differences between the EFG implementations are the same as those for the comparison of main group diatomics in Table 2, apart from the spin–orbit corrections which were consistently used in the present work. There is overall good agreement with the reference data of ref 2, indicating that spin–orbit relativistic contributions to the Nb and Ru EFGs are not very important. However, the calculations for this work were carried out at the spin–orbit level to demonstrate that routine calculations of this type can be performed for these and other metal complexes and organometallic systems. For comparison, we have also carried out calculations on the Nb complexes with the TANH basis, which yields very similar results. The differences from the ANO basis are smaller than 0.02 au in magnitude, which is significantly smaller than the variations among the different functionals.

The origin of a nonzero EFG in the Ru complexes has been analyzed recently.² For [Ru(bipy₃)]²⁺, the origin of the relatively small field gradient is a compression of the pseudo-octahedral nitrogen arrangement along the 3-fold symmetry axis of the complex, causing less involvement of the Ru 4d_{z²} orbital in metal–ligand π backbonding than for the other two occupied Ru 4d orbitals (negative V_{33} contributions), which is only partially counterbalanced by positive V_{33} contributions from Ru–N σ bonding orbitals created by the structural distortion. In ruthenocene, the EFG is caused by a strongly nonspherical charge distribution in the 4d shell due to the formal (4d_{xy})²(4d_{x²−y²)²(4d_{z²})² configuration, which is to some extent balanced by significant π donation from}

Cp to the metal and by δ backbonding from the metal. In [RuCl₄N][−], the large EFG is mainly caused by the double occupation of just one 4d orbital, creating a strongly nonspherical electronic charge density in the Ru 4d shell.

The calculations performed here with different density functionals show a modest dependence of the EFGs on the description of exchange and correlation, with a trend of more positive Ru EFG along the series revPBE, B3LYP, and CAM-B3LYP. This trend leads to slightly better agreement with experimentally derived EFGs for the tris-bipyridyl complex with CAM-B3LYP, but for [RuCl₄N][−] the EFG calculated with CAM-B3LYP appears to be somewhat overestimated. Overall, however, the functional dependence is not as pronounced as those for the transition metal diatomics of the previous section. For instance, unlike for the case of Au and Cu, the differences between functionals are much smaller than the range of EFGs spanned by the set of complexes (which, of course, depends on the selection of the test set). The performance of all functionals and the combination of basis sets used here with Z4 therefore appears to be suitable for routine calculations on these larger metal complexes to rationalize and analyze EFG trends among them.

The gas-phase calculated EFGs in the Nb complexes show greater discrepancy with experimental values than the Ru EFGs, likely because the experimental results are derived from solid-state NMR measurements, while longer range effects on the EFG from the crystal environment and effects from nearest neighbor contacts are not included in the calculations. The origin of the EFGs in the Nb complexes can be traced back to a complex interplay of metal–ligand bonding and backbonding involving the Nb 4d shell, which is sensitive to the computational model used to describe the environment of the complexes. As for the Ru complexes, the EFGs calculated for this work are in good agreement with the SR-Z4 STO basis set data from ref 2, and the level of theory appears to be well suited for routine calculations at the two-component relativistic DFT level of theory.

4.4. Uranyl Salts. Uranyl, UO₂²⁺, ions are important benchmark systems for actinide electronic structure calculations due to their size, symmetric geometry, closed shell nature, chemical stability, and presence in plenty of compounds. As deceptively simple that uranyl might appear, it challenges quantum mechanical approaches by requiring high level electron–electron correlation methods, flexible basis sets, and—of course—relativistic methods (preferably including spin–orbit coupling).^{94–96} Since the EFG is a

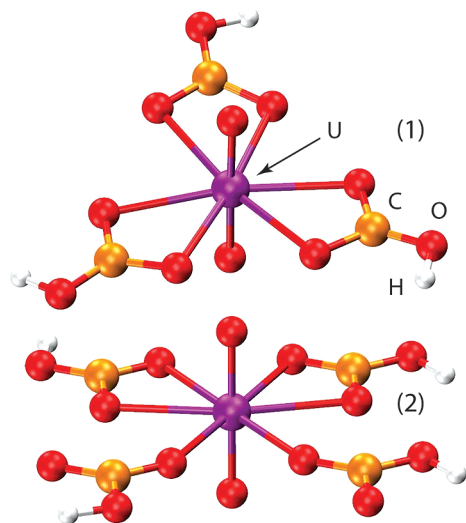


Figure 6. Geometries of uranium complexes. **1** has a U–O bond length of 1.791 Å and a O–U–O bond angle of 179.0°. **2** has a U–O bond length of 1.743 Å and a O–U–O bond angle of 178.6°.

Table 9. Computed Uranium V_{33} , in Atomic Units, in Uranyl and Uranyl Complexes.^a Spin–Orbit Z4 Computations

	CAM-B3LYP-A	B3LYP	revPBE	BP
[UO ₂ (HCO ₃) ₃] ⁻ (1) ^c	9.02	7.36	5.76	
[UO ₂] ²⁺ of (1)	-7.99	-9.46	-10.96	
[UO ₂ (HCO ₃) ₄] ²⁻ (2) ^c	11.19	9.40	7.66	
[UO ₂] ²⁺ of (2)	-5.12	-6.67	-8.36	-8.88
[UO ₂] ²⁺ of (2) ^b	-5.32	-6.86	-8.58	-9.12
[UO ₂] ²⁺ (3)				-11.08 (-10.58)

^a Using the ANO-RCC basis set for uranium and TZVPP for oxygens and using Cartesian (6d, 10f, 15 g) angular functions unless explicitly stated otherwise. ^b Using the ANO-RCC basis for U and O. ^c Using spherical (5d,7f,9g) angular functions.

sensitive indicator of the ground state electronic structure around a metal, it serves as a particularly suitable “probe” for comparing different density functionals. A broader interest in studying uranium and other actinide compounds arises from the nuclear waste problem. In particular, carbonate complexes have been identified as important uranyl species in natural waters.⁹⁷

Figure 6 shows the geometries of two uranyl carbonate complexes considered here for calculations of uranium EFGs. The geometries were taken from X-ray crystalline structures of two uranyl salts:⁹⁵ (NH₄)₄UO₂(CO₃)₃ (**1**) and rutherfordine (UO₂(CO₃); **2**). In these salts, the uranyl moiety has different coordinations by the carbonate ligands. In some calculations, protons were added to the clusters cut from the crystal structures, changing CO₃²⁻ to HCO₃⁻, in order to reduce the overall charges. For comparison with ref 48, we included in our study a third geometry labeled **3**, which is a perfectly linear UO₂ molecule having a U–O bond length of 1.78 Å. In Table 9, the uranium EFGs are collected for complexes **1** and **2**, calculated with the CAM-B3LYP-A, B3LYP, and revPBE functionals. Additional EFG calculations were performed on the corresponding uranyl moiety in **1** and **2** using OUO bond angles and U–O bond lengths as adopted in geometries **1** and **2**, respectively.

From the results shown in the table, we observe a change in sign in the EFG at U upon complexation with carbonate, from negative in free uranyl and for the isolated uranyl moieties of **1** and **2** to positive in the presence of equatorial ligands. Such a behavior of the uranium EFGs due to ligand coordination has been discussed in previous works by Belanzoni et al. and de Jong et al.^{48,98} and has been studied for different coordinations. Table 9 also lists the EFG at U for geometry **3** calculated with the BP functional. This value is in very good agreement with the EFG shown in parentheses which was taken from Belanzoni et al.⁴⁸ (STO basis sets, same geometry). The table also demonstrates the effect of using different sized basis sets for the uranyl oxygens: TZVPP (11s,6p,2d,1f) versus an uncontracted ANO-RCC basis (14s,9p,4d,3f). We observe that the EFG at uranium is increased by about 0.2 au with the larger basis set with the different functionals. Compared to the more than 3 au variation among the functionals, this basis set effect is not particularly significant, and the more economic TZVPP basis can be used in calculations of the metal EFG. Additional SR-Z4 calculations discussed below indicated that even with the relatively small 6-31G* basis for O, the uranium EFG does not change significantly. The magnitudes of the EFGs at U in the free uranyl ion for the four exchange-correlation functionals tend to decrease in the order shown in the table from right to left, i.e., when going from nonhybrid to a standard hybrid to the Coulomb-attenuated hybrid. The trend for the complexated uranyl ions **1** and **2** goes in the same direction of less negative EFG, thereby increasing the magnitude of the positive EFG from BP and revPBE to CAM-B3LYP. Experimental EFGs have to our knowledge not been determined for the carbonate complexes. However, for the related nitrate complexes investigated by Belanzoni et al., the experimental estimate derived from Mössbauer data is $+8.38 \pm 0.13$ au,⁹⁹ which is in reasonable agreement with our data.

It is interesting to dissect the uranyl EFG and the sign change upon complexation in the equatorial plane with the localized orbital EFG analysis. To this end, we have performed such calculations at the SR-Z4 level of theory, similar to those for Cu and Au of section 4.2. A thorough analysis of the uranyl EFG based on MOs and AOs has been carried out by Belanzoni et al.,⁴⁸ which therefore allows comparisons with the results from the previous analysis. See also de Jong et al.⁹⁸ Of particular interest are the AO contributions from the U 6p, 6d, and 5f. It was pointed out by Pyykkö that the positive EFG in the experimentally observed systems might be a signature of a 6p hole.^{100,101} Belanzoni et al. confirmed the presence of a partial 6p hole, but as in our study, the net EFG in free uranyl was found to be negative in part due to a nonspherical electron distribution in the valence 5f shell from the bonding of uranium to the oxygens. Only in the systems with equatorial ligands is a positive EFG computed.

The relative magnitudes of EFG contributions from atomic shells were already discussed in section 4.2 (see also ref 2). For holes in otherwise filled atomic shells, the “contributions” (i.e., lack of electronic EFG) have the opposite sign. For uranyl oriented along the z axis, V_{33} is equal to V_{zz} . The f

orbitals in Figure 5 are then conveniently grouped into a single f_σ , two f_π , two f_δ , and two f_ϕ . The f_δ 's do not contribute to the EFG. On the basis of their computational analysis, Belanzoni et al. assigned to free UO_2^{2+} an approximate effective electron configuration of $6s^2 6p^6 5f_\sigma^1 5f_\pi^2 6d_\pi^1 7s^2$ for U. The effective atomic population can be considered formally as arising from the U(VI) f^0 ion binding to two O^{2-} 's. From Figure 5, it is seen that the EFG from the 5f occupations would be cumulative since f_σ and f_π contribute with negative signs to V_{zz} . Partial d_π occupations would cause a negative V_{33} as well, while spherical 6s, 7s, and the filled 6p shells would not be expected to give rise to a field gradient at the U nucleus. The main source of the large negative EFG in free UO_2^{2+} would then be from the nonspherical 5f charge density. Upon coordination by ligands in the equatorial plane, the formally empty f_ϕ orbitals become suitable electronic charge acceptors (see Figure 5). The potentially large positive V_{33} contributions from these and partially occupied equatorial 6d orbitals on uranium were identified in ref 48 as a major source for the EFG sign change. However, it was also noted that it was "difficult to interpret the total EFG quantitatively in terms of a simple chemical bonding picture", because of canceling terms that individually far exceeded the magnitude of the net EFG.

When considering the effective uranium atomic configuration as devised by Belanzoni et al., relative to U(VI) there are six excess electrons at U, so the bonding pattern might be considered as $[\text{O}\equiv\text{U}\equiv\text{O}]^{2+}$. This is in fact the calculated bonding pattern assigned from the NBO analysis based on the calculated scalar ZORA density matrix, along with an assignment of $5f^{2.50} 6s^2 6p^2 6d^{0.45} 7s^{0.06}$ for the uranium valence configuration, based on its natural atomic orbital populations. The d and f occupations are quite consistent with the effective populations deduced by Belanzoni et al. A low occupation of one of each of the U–O bonding NBOs (1.86 for each σ bond NBO) indicates that these are not to be interpreted as simple textbook covalent bonds but partially as delocalized lone pairs.

The EFG data for the analysis are collected in Table 10. The sign change upon complexation in the equatorial plane is obtained also at the scalar relativistic level. A comparison with the spin–orbit data of Table 9 indicates fairly pronounced spin–orbit effects, but for semiquantitative purposes, scalar relativistic calculations are certainly suitable. The analysis data were obtained with the 6-31G* basis for the ligand atoms instead of TZVPP because of difficulties that the NBO procedure had assigning the bonding patterns in complex **1** when using the larger ligand basis. For comparison, the SR-Z4 EFG in uranyl calculated with TZVPP for the oxygens is -7.422 and is seen to be very close to the value listed in Table 10. According to the NLMO decomposition of the field gradient in UO_2^{2+} , the U–O bonds are the source of a large negative EFG. A breakdown of the σ and π bonding NLMOs shows virtually no U p character and a strong dominance of f_σ in the σ bonds, along with about 78% uranium f_π and 20% d_π at U for the π bonding NLMOs. Because of the nature of the participating U orbitals, donation of electronic charge from the oxygens into f_σ , f_π , d_σ , and d_π naturally yields a negative EFG from

Table 10. NLMO Analysis of V_{33} (au) at Uranium in the UO_2^{2+} Moiety of **1** and the Uranyl Salt (**1**)^a

NLMO	$[\text{UO}_2]^{2+}$	1
$\sigma(\text{U}-\text{O})^b$ (ea., $\times 2$)	-2.641	-2.372
$\pi(\text{U}-\text{O})^c$ (ea., $\times 4$)	-0.710	-0.432
eq. U–O bonds		2.529
Σ U–O bonding	-8.122	-3.943
O core ax.	-0.176	-0.202
O LP ^d ax.	-2.684	-2.226
Σ O-eq LP		0.092
U core ^e	-1.239	1.747
U 6s	-1.272	-0.945
U 6p _σ	-66.830	-104.239
U 6p _π (ea.)	36.459	57.945
Σ U 6s, 6p	4.815	10.706
O core eq.		0.154
Σ other NLMOs		0.485
total calcd	-7.406	6.815

^a SRZ4, B3LYP. See text for details. ^b $\sigma(\text{U}-\text{O})$ (av.): 30.4% U [s (2.7%), p (0.2%), d (8.5%), f (88.6%)], 67.4% O_α [s (6.4%), p (93.2%), d (0.4%), f (0.0%)], 2.2% O_β [s (11.8%), p (88.0%), d (0.2%), f (0.0%)]. ^c $\pi(\text{U}-\text{O})$ (av.): 20.5% U [s (0.0%), p (0.3%), d (20.1%), f (78.3%), g (1.4%)], 78.9% O_α [s (0.0%), p (99.7%), d (0.3%), f (0.0%), g (0.0%)], 0.7% O_β [s (0.1%), p (96.0%), d (3.9%), f (0.0%), g (0.0%)]. ^d O LP (ea.): 0.6% U [s (83.1%), p (1.4%), d (2.3%), f (12.8%), g (0.5%)], 99.4% O_α [s (93.5%), p (6.5%), d (0.0%), f (0.0%), g (0.0%)]. ^e Core orbitals: 1s–5d.

the U–O bonds. The second largest influence on the EFG comes from the formally filled 6sp semicore shell. The 6p contribution is indeed positive, as anticipated, and partially balanced by a significant negative contribution from U "6s" (which is obviously not perfectly spherically symmetric). The cumulative U core (1s through 5d) contributions to the EFG are also negative. Somewhat unintuitive are additional sizable negative contributions from the oxygen σ lone pair NLMOs. However, the NLMO breakdown in Table 10 indicates that there is some delocalization of these orbitals in UO_2^{2+} , with participation of U 7s and 5f_σ, with the resulting partial occupation of the latter again giving rise to a negative EFG. One should also keep in mind that the NLMO decomposition includes electronic and nuclear terms, that is, EFG contributions related to how well or not a particular orbital shields the charge from a neighboring nucleus. The localized orbital decomposition of the uranyl EFG is therefore seen to be rather straightforward, it yields a compact analysis, and overall it supports the conclusions drawn by Belanzoni et al.⁴⁸

For the carbonate complex **1**, the juxtaposition of the analysis data with those for UO_2^{2+} reveals several trends. The equatorial U–O interactions show up directly in the form of positive contributions to the U EFG, consistent with an involvement of 6p_δ and 5f_φ (see also Figure 5). Further, the electron density rearrangement caused by the presence of the equatorial ligands leads to a reduced EFG from the axial U–O bonds (both σ and π) relative to free UO_2^{2+} , to a sign change of the U core EFG, and to a significantly increased positive EFG from the uranium 6p shell. The almost 11 au EFG contributions from the 6p shell now represent the largest individual influence on the EFG. Closer inspection of the NBOs revealed an occupation of only 1.78 for 6p_σ in

complex **1**, which very clearly indicates a partial $6p_{\sigma}$ “hole” and rationalizes the overall positive EFG in the complex. The corresponding NLMO has almost 5% contribution from the axial oxygens. For comparison, the $6p_{\pi}$ NBOs have occupations greater than 1.99.

5. Concluding Remarks and Outlook

The new implementation of the Z4 and SR-Z4 approaches for calculations of nuclear EFGs in molecules has demonstrated good agreement with published benchmark data where available, and with experimental results. For the main group diatomics previously studied by van Lenthe and Baerends with a nonhybrid Z4 method,¹³ a comparison of different functionals including a range-separated hybrid has shown that the EFGs are quite robust with respect to the choice of functional. Modest improvements are obtained with the hybrid functionals. For transition metals, we were able to confirm the strong dependence on the functional in Cu and Au diatomic molecules, but for larger transition metal complexes the results did not vary so dramatically. Relativistic effects for heavy nucleus EFGs are highly important, with the scalar relativistic effects on the two-component density providing the largest effect. The Z4 correction was in many cases of similar magnitude as effects on the EFG due to spin–orbit coupling. Both should be considered for accurate computations. The basis set combinations used in this work allow for comparatively fast routine EFG calculations. With DFT, the high angular momentum functions in the ANO basis sets appear to be of little benefit in EFG computations on transition metal systems. A localized MO analysis was implemented to assist chemically intuitive interpretations of the results. For CuF, CuCl, AuCl, and the uranyl systems, the localized orbital decomposition has clearly shown the role of individual valence metal p, d, and f shells; bonding and nonbonding orbitals; and core orbitals. The NLMO decomposition leads to rather compact analyses. For f orbitals, a Townes-Dailey like AO model has been set up. The calculations on AuCl showed that for a relatively unperturbed f core shell the EFG signs and magnitudes of the f_{σ} , f_{π} , f_{δ} , and f_{ϕ} orbitals almost exactly match the model. For uranyl, the analysis demonstrated the delicate balance between EFG contributions from the various semicore and valence orbitals, leading to conclusions that agree well with those drawn by Belanzoni et al.⁴⁸

Acknowledgment. The authors acknowledge support of this research from the Center of Computational Research at SUNY Buffalo, and financial support from the US Department of Energy, grant no. DE-SC0001136 (BES Heavy Element Chemistry Program). Some calculations were performed using EMSL, a national scientific user facility sponsored by the Department of Energy’s Office of Biological and Environmental Research and located at Pacific Northwest National Laboratory. N.G. would like to acknowledge the DOE BES Heavy Element Chemistry Program (PI: De Jong) of the U.S. Department of Energy, Office of Science for helping support the ZORA code development.

Appendix

For completeness, numerical data for diatomics calculated with different functionals are provided in Tables 11–15).

Table 11. Calculated Electric Field Gradient, V_{33} , at the Halide Nucleus for a Set of Diatomics, Using the BLYP Functional

		NR	SR-Z4	Z4	observed
AlCl	³⁵ Cl	0.3994	0.4028	0.4030	0.4602 ^a
GaCl	³⁵ Cl	0.6233	0.6312	0.6330	0.6880 ^b
InCl	³⁵ Cl	0.6407	0.6603	0.6724	0.6933 ^b
CuCl	³⁵ Cl	2.3592	2.5695	2.5698	1.675 ^a
HCl	³⁵ Cl	3.6209	3.6657	3.6661	3.516 ^c
ICI	³⁵ Cl	4.6109	4.6972	4.5785	4.472 ^c
BrCl	³⁵ Cl	5.4136	5.4758	5.4553	5.336 ^c
AlBr	⁷⁹ Br	0.9096	0.9734	0.9773	1.112 ^d
GaBr	⁷⁹ Br	1.2458	1.3254	1.3344	1.50 ^b
InBr	⁷⁹ Br	1.2954	1.3940	1.4274	1.57 ^b
CuBr	⁷⁹ Br	4.6492	5.3150	5.3196	3.71 ^e
HBr	⁷⁹ Br	7.1944	7.6538	7.6804	7.55 ^c
IBr	⁷⁹ Br	9.3754	9.9780	9.8329	9.89 ^b
BrCl	⁷⁹ Br	11.6868	12.3605	12.3422	12.4 ^c
All	¹²⁷ I	1.3886	1.6377	1.6754	1.90 ^b
Gal	¹²⁷ I	1.6771	1.9495	1.9984	2.28 ^b
InI	¹²⁷ I	1.7646	2.0604	2.1479	2.38 ^b
TII	¹²⁷ I	1.7660	2.0472	2.5754	2.70 ^b
Cul	¹²⁷ I	6.4486	7.9596	7.9937	5.79 ^e
HI	¹²⁷ I	9.7557	11.2543	11.4814	11.3 ^c
IBr	¹²⁷ I	14.8109	16.8499	16.7877	16.8 ^c
ICI	¹²⁷ I	15.8279	18.0085	17.8228	18.1 ^c
IF	¹²⁷ I	18.6571	21.1839	20.8508	21.2 ^c
AlF	²⁷ Al	-1.1388	-1.1432	-1.1433	-1.096 ^b
AlCl	²⁷ Al	-0.9400	-0.9420	-0.9421	-0.8828 ^b
AlBr	²⁷ Al	-0.8738	-0.8699	-0.8692	-0.8130 ^d
All	²⁷ Al	-0.8165	-0.8038	-0.7997	-0.7417 ^b
GaF	⁶⁹ Ga	-2.5865	-2.6624	-2.6714	-2.76 ^b
GaCl	⁶⁹ Ga	-2.2630	-2.3287	-2.3371	-2.38 ^f
GaBr	⁶⁹ Ga	-2.1457	-2.1981	-2.2043	-2.24 ^b
Gal	⁶⁹ Ga	-2.0441	-2.0729	-2.0702	-2.10 ^b
InF	¹¹⁵ In	-3.6863	-3.9757	-4.0698	-4.18 ^b
InCl	¹¹⁵ In	-3.3674	-3.6310	-3.7216	-3.79 ^b
InBr	¹¹⁵ In	-3.2481	-3.4905	-3.5759	-3.65 ^b
InI	¹¹⁵ In	-3.1454	-3.3500	-3.4157	-3.49 ^b

^a Ref 65. ^b Ref 66. ^c Ref 67. ^d Ref 68. ^e Ref 69. ^f Ref 70.

Table 12. Calculated Electric Field Gradient, V_{33} , at the Halide Nucleus for a Set of Diatomics, Using the the B3LYP Functional

		NR	SR-Z4	Z4	observed
AlCl	³⁵ Cl	0.4387	0.4425	0.4426	0.4602 ^a
GaCl	³⁵ Cl	0.6727	0.6815	0.6834	0.6880 ^b
InCl	³⁵ Cl	0.6867	0.7083	0.7201	0.6933 ^b
CuCl	³⁵ Cl	2.0518	2.2554	2.2558	1.675 ^a
HCl	³⁵ Cl	3.6204	3.6655	3.6660	3.516 ^c
ICI	³⁵ Cl	4.6494	4.7359	4.6182	4.472 ^c
BrCl	³⁵ Cl	5.4777	5.5405	5.5204	5.336 ^c
AlBr	⁷⁹ Br	0.9873	1.0571	1.0614	1.112 ^d
GaBr	⁷⁹ Br	1.3459	1.4339	1.4434	1.50 ^b
InBr	⁷⁹ Br	1.3894	1.4988	1.5318	1.57 ^b
CuBr	⁷⁹ Br	4.0692	4.7062	4.7176	3.71 ^e
HBr	⁷⁹ Br	7.1945	7.6624	7.6899	7.55 ^c
IBr	⁷⁹ Br	9.4886	10.1104	9.9667	9.89 ^b
BrCl	⁷⁹ Br	11.8773	12.5782	12.5671	12.4 ^c
All	¹²⁷ I	1.5135	1.7884	1.8303	1.90 ^b
Gal	¹²⁷ I	1.8355	2.1397	2.1934	2.28 ^b
InI	¹²⁷ I	1.9136	2.2425	2.3387	2.38 ^b
TII	¹²⁷ I	1.9085	2.2287	2.5395	2.70 ^b
Cul	¹²⁷ I	5.6985	7.1362	7.2312	5.79 ^e
HI	¹²⁷ I	9.8002	11.3443	11.5819	11.3 ^c

Table 12. Continued

		NR	SR-Z4	Z4	observed
IBr	¹²⁷ I	15.1145	17.2600	17.2468	16.8 ^c
ICI	¹²⁷ I	16.1665	18.4638	18.3318	18.1 ^c
IF	¹²⁷ I	19.1211	21.7892	21.5091	21.2 ^c
AIF	²⁷ Al	-1.1600	-1.1646	-1.1646	-1.096 ^b
AlCl	²⁷ Al	-0.9576	-0.9597	-0.9597	-0.8828 ^b
AlBr	²⁷ Al	-0.8909	-0.8868	-0.8862	-0.8130 ^d
All	²⁷ Al	-0.8341	-0.8207	-0.8171	-0.7417 ^b
GaF	⁶⁹ Ga	-2.6583	-2.7370	-2.7464	-2.76 ^b
GaCl	⁶⁹ Ga	-2.3211	-2.3882	-2.3968	-2.38 ^f
GaBr	⁶⁹ Ga	-2.2002	-2.2524	-2.2589	-2.24 ^b
Gal	⁶⁹ Ga	-2.1007	-2.1282	-2.1269	-2.10 ^b
InF	¹¹⁵ In	-3.8055	-4.1081	-4.2062	-4.18 ^b
InCl	¹¹⁵ In	-3.4756	-3.7499	-3.8447	-3.79 ^b
InBr	¹¹⁵ In	-3.3548	-3.6062	-3.6963	-3.65 ^b
InI	¹¹⁵ In	-3.2588	-3.4704	-3.5452	-3.49 ^b

^a Ref 65. ^b Ref 66. ^c Ref 67. ^d Ref 68. ^e Ref 69. ^f Ref 70.**Table 13.** Calculated Electric Field Gradient, V_{33} , at the Halide Nucleus for a Set of Diatomics, Using the CAM-B3LYP-A Functional

		NR	SR-Z4	Z4	observed
AlCl	³⁵ Cl	0.4482	0.4518	0.4519	0.4602 ^a
GaCl	³⁵ Cl	0.6934	0.7013	0.7019	0.6880 ^b
InCl	³⁵ Cl	0.7064	0.7260	0.7313	0.6933 ^b
CuCl	³⁵ Cl	1.8731	2.0695	2.0699	1.675 ^a
HCl	³⁵ Cl	3.6148	3.6603	3.6610	3.516 ^c
ICI	³⁵ Cl	4.6548	4.7355	4.7395	4.472 ^c
BrCl	³⁵ Cl	5.5104	5.5716	5.5730	5.336 ^c
AlBr	⁷⁹ Br	1.0249	1.0952	1.1013	1.112 ^d
GaBr	⁷⁹ Br	1.4117	1.5007	1.5100	1.50 ^b
InBr	⁷⁹ Br	1.4535	1.5625	1.5809	1.57 ^b
CuBr	⁷⁹ Br	3.7383	4.3511	4.3777	3.71 ^e
HBr	⁷⁹ Br	7.2244	7.6984	7.7394	7.55 ^c
IBr	⁷⁹ Br	9.5934	10.2194	10.2786	9.89 ^b
BrCl	⁷⁹ Br	12.0816	12.8040	12.8700	12.4 ^c
All	¹²⁷ I	1.6217	1.9082	1.9727	1.90 ^b
Gal	¹²⁷ I	1.9908	2.3116	2.3890	2.28 ^b
InI	¹²⁷ I	2.0631	2.4077	2.4989	2.38 ^b
TII	¹²⁷ I	2.0566	2.3743	2.6097	2.70 ^b
CuI	¹²⁷ I	5.2757	6.6775	6.8936	5.79 ^e
HI	¹²⁷ I	9.9227	11.5014	11.8735	11.3 ^c
IBr	¹²⁷ I	15.4910	17.7226	18.2879	16.8 ^c
ICI	¹²⁷ I	16.5887	18.9846	19.5867	18.1 ^c
IF	¹²⁷ I	19.6142	22.3860	23.1013	21.2 ^c
AIF	²⁷ Al	-1.1764	-1.1810	-1.1810	-1.096 ^b
AlCl	²⁷ Al	-0.9828	-0.9849	-0.9849	-0.8828 ^b
AlBr	²⁷ Al	-0.9170	-0.9124	-0.9123	-0.8130 ^d
All	²⁷ Al	-0.8640	-0.8492	-0.8484	-0.7417 ^b
GaF	⁶⁹ Ga	-2.7144	-2.7939	-2.8038	-2.76 ^b
GaCl	⁶⁹ Ga	-2.3988	-2.4662	-2.4755	-2.38 ^f
GaBr	⁶⁹ Ga	-2.2802	-2.3313	-2.3399	-2.24 ^b
Gal	⁶⁹ Ga	-2.1921	-2.2160	-2.2229	-2.10 ^b
InF	¹¹⁵ In	-3.9160	-4.2220	-4.3274	-4.18 ^b
InCl	¹¹⁵ In	-3.6177	-3.8945	-3.9977	-3.79 ^b
InBr	¹¹⁵ In	-3.5026	-3.7544	-3.8555	-3.65 ^b
InI	¹¹⁵ In	-3.4274	-3.6360	-3.7339	-3.49 ^b

^a Ref 65. ^b Ref 66. ^c Ref 67. ^d Ref 68. ^e Ref 69. ^f Ref 70.**Table 14.** Calculated Electric Field Gradient, V_{33} , at the Halide Nucleus for a Set of Diatomics, Using the CAM-B3LYP-B Functional

		NR	SR-Z4	Z4	observed
AlCl	³⁵ Cl	0.4555	0.4589	0.4590	0.4602 ^a
GaCl	³⁵ Cl	0.7111	0.7183	0.7192	0.6880 ^b
InCl	³⁵ Cl	0.7255	0.7434	0.7488	0.6933 ^b
CuCl	³⁵ Cl	1.7485	1.9368	1.9371	1.675 ^a
HCl	³⁵ Cl	3.5972	3.6426	3.6434	3.516 ^c

Table 14. Continued

		NR	SR-Z4	Z4	observed
ICI	³⁵ Cl	4.6628	4.7393	4.7446	4.472 ^c
BrCl	³⁵ Cl	5.5378	5.5978	5.5996	5.336 ^c
AlBr	⁷⁹ Br	1.0521	1.1228	1.1293	1.112 ^d
GaBr	⁷⁹ Br	1.4638	1.5538	1.5636	1.50 ^b
InBr	⁷⁹ Br	1.5083	1.6169	1.6359	1.57 ^b
CuBr	⁷⁹ Br	3.5011	4.0911	4.1165	3.71 ^e
HBr	⁷⁹ Br	7.2211	7.6980	7.7391	7.55 ^c
IBr	⁷⁹ Br	9.6806	10.3109	10.3737	9.89 ^b
BrCl	⁷⁹ Br	12.2407	12.9805	13.0478	12.4 ^c
All	¹²⁷ I	1.6956	1.9907	2.0578	1.90 ^b
Gal	¹²⁷ I	2.1060	2.4399	2.5217	2.28 ^b
InI	¹²⁷ I	2.1794	2.5355	2.6319	2.38 ^b
TII	¹²⁷ I	2.1736	2.4872	2.7337	2.70 ^b
CuI	¹²⁷ I	4.9615	6.3195	6.5247	5.79 ^e
HI	¹²⁷ I	9.9852	11.5851	11.9611	11.3 ^c
IBr	¹²⁷ I	15.7956	18.0980	18.6797	16.8 ^c
ICI	¹²⁷ I	16.9230	19.3979	20.0187	18.1 ^c
IF	¹²⁷ I	20.0120	22.8680	23.6057	21.2 ^c
AIF	²⁷ Al	-1.1870	-1.1916	-1.1917	-1.096 ^b
AlCl	²⁷ Al	-0.9994	-1.0014	-1.0015	-0.8828 ^b
AlBr	²⁷ Al	-0.9335	-0.9286	-0.9285	-0.8130 ^d
All	²⁷ Al	-0.8821	-0.8662	-0.8654	-0.7417 ^b
GaF	⁶⁹ Ga	-2.7589	-2.8390	-2.8492	-2.76 ^b
GaCl	⁶⁹ Ga	-2.4573	-2.5249	-2.5344	-2.38 ^f
GaBr	⁶⁹ Ga	-2.3381	-2.3882	-2.3970	-2.24 ^b
Gal	⁶⁹ Ga	-2.2549	-2.2759	-2.2828	-2.10 ^b
InF	¹¹⁵ In	-4.0069	-4.3162	-4.4245	-4.18 ^b
InCl	¹¹⁵ In	-3.7303	-4.0091	-4.1156	-3.79 ^b
InBr	¹¹⁵ In	-3.6164	-3.8681	-3.9724	-3.65 ^b
InI	¹¹⁵ In	-3.5525	-3.7573	-3.8582	-3.49 ^b

^a Ref 65. ^b Ref 66. ^c Ref 67. ^d Ref 68. ^e Ref 69. ^f Ref 70.**Table 15.** Calculated Electric Field Gradient, V_{33} , at the Halide Nucleus for a Set of Diatomics, Using the CAM-B3LYP* Functional

		NR	SR-Z4	Z4	observed
AlCl	³⁵ Cl	0.4972	0.5018	0.5020	0.4602 ^a
GaCl	³⁵ Cl	0.7389	0.7485	0.7491	0.6880 ^b
InCl	³⁵ Cl	0.7367	0.7598	0.7651	0.6933 ^b
CuCl	³⁵ Cl	1.5905	1.7706	1.7710	1.675 ^a
HCl	³⁵ Cl	3.6640	3.7104	3.7112	3.516 ^c
ICI	³⁵ Cl	4.7001	4.7869	4.7918	4.472 ^c
BrCl	³⁵ Cl	5.5935	5.6574	5.6592	5.336 ^c
AlBr	⁷⁹ Br	1.1043	1.1836	1.1902	1.112 ^d
GaBr	⁷⁹ Br	1.4838	1.5828	1.5926	1.50 ^b
InBr	⁷⁹ Br	1.4984	1.6196	1.6384	1.57 ^b
CuBr	⁷⁹ Br	3.1880	3.7451	3.7601	3.71 ^e
HBr	⁷⁹ Br	7.2972	7.7848	7.8264	7.55 ^c
IBr	⁷⁹ Br	9.6872	10.3426	10.4046	9.89 ^b
BrCl	⁷⁹ Br	12.2806	13.0320	13.0995	12.4 ^c
All	¹²⁷ I	1.6993	2.0139	2.0822	1.90 ^b
Gal	¹²⁷ I	2.0574	2.4062	2.4873	2.28 ^b
InI	¹²⁷ I	2.0975	2.4720	2.5656	2.38 ^b
TII	¹²⁷ I	2.0759	2.4287	2.6726	2.70 ^b
CuI	¹²⁷ I	4.5304	5.7966	5.9847	5.79 ^e
HI	¹²⁷ I	10.0133	11.6495	12.0283	11.3 ^c
IBr	¹²⁷ I	15.7276	18.0668	18.6463	16.8 ^c
ICI	¹²⁷ I	16.8661	19.3791	19.9979	18.1 ^c
IF	¹²⁷ I	20.0830	23.0210	23.7621	21.2 ^c
AIF	²⁷ Al	-1.2001	-1.2048	-1.2049	-1.096 ^b
AlCl	²⁷ Al	-0.9925	-0.9946	-0.9947	-0.8828 ^b
AlBr	²⁷ Al	-0.9240	-0.9195	-0.9194	-0.8130 ^d

Table 15. Continued

		NR	SR-Z4	Z4	observed
All	²⁷ Al	-0.8671	-0.8520	-0.8512	-0.7417 ^b
GaF	⁶⁹ Ga	-2.7718	-2.8543	-2.8644	-2.76 ^b
GaCl	⁶⁹ Ga	-2.4221	-2.4903	-2.4995	-2.38 ^f
GaBr	⁶⁹ Ga	-2.2949	-2.3461	-2.3547	-2.24 ^b
GaI	⁶⁹ Ga	-2.1969	-2.2199	-2.2266	-2.10 ^b
InF	¹¹⁵ In	-3.9819	-4.3001	-4.4076	-4.18 ^b
InCl	¹¹⁵ In	-3.6523	-3.9393	-4.0438	-3.79 ^b
InBr	¹¹⁵ In	-3.5292	-3.7895	-3.8914	-3.65 ^b
InI	¹¹⁵ In	-3.4438	-3.6586	-3.7568	-3.49 ^b

^a Ref 65. ^b Ref 66. ^c Ref 67. ^d Ref 68. ^e Ref 69. ^f Ref 70.

Note Added after ASAP Publication. This article was published ASAP on August 10, 2010. The title of Table I has been modified. The correct version was published on August 19, 2010.

References

- (1) Lucken, E. A. C. *Nuclear quadrupole coupling constants*; Academic Press: New York, 1969.
- (2) Autschbach, J.; Zheng, S.; Schurko, R. W. *Concepts Magn. Reson. A* **2010**, *36A*, 84–126.
- (3) Das, T. P. *Nuclear quadrupole resonance spectroscopy*; Academic Press: New York, 1958.
- (4) Dillon, K. B. Nuclear quadrupole resonance spectroscopy. In *Spectroscopic properties of inorganic and organometallic compounds*; The Royal Society of Chemistry: London, 2005; Vol. 37.
- (5) Schwerdtfeger, P.; Söhnel, T.; Pernpointner, M.; Laerdahl, J. K.; Wagner, F. E. *J. Chem. Phys.* **2001**, *115*, 5913–5924.
- (6) Barone, G.; Mastalerz, G.; Reiher, M.; Lindh, R. *J. Phys. Chem. A* **2008**, *112*, 1666–1672.
- (7) Kowalewski, J.; Mäler, L. *Nuclear spin relaxation in liquids: Theory, experiments, and applications*; Taylor & Francis: New York, 2006.
- (8) Bryce, D. L.; Eichele, K.; Wasylishen, R. E. *Inorg. Chem.* **2003**, *42*, 5085–5096.
- (9) Wong, A.; Pike, K. J.; Jenkins, R.; Clarkson, G. J.; Anupold, T.; Howes, A. P.; Crout, D. H. G.; Samoson, A.; Dupree, R.; Smith, M. E. *J. Phys. Chem. A* **2006**, *110*, 1824–1835.
- (10) Bryce, D. L.; Sward, G. D. *Magn. Reson. Chem.* **2006**, *44*, 409–450.
- (11) Pernpointner, M.; Visscher, L. *J. Chem. Phys.* **2001**, *114*, 10389.
- (12) Visscher, L.; Enevoldsen, T.; Saue, T.; Oddershede, J. *J. Chem. Phys.* **1998**, *109*, 9677–9684.
- (13) van Lenthe, E.; Baerends, E. J. *J. Chem. Phys.* **2000**, *112*, 8279–8292.
- (14) Neese, F.; Wolf, A.; Fleig, T.; Reiher, M.; Hess, B. A. *J. Chem. Phys.* **2005**, *122*, 204107.
- (15) Mastalerz, R.; Barone, G.; Lindh, R.; Reiher, M. *J. Chem. Phys.* **2007**, *127*, 074105.
- (16) Rutkowski, A. *J. Phys. B* **1986**, *19*, 149–158.
- (17) Rutkowski, A.; Schwarz, W. H. E. *Theor. Chim. Acta* **1990**, *76*, 391–410.
- (18) Cazzolia, G.; Puzzarina, C.; Stopkowicz, S.; Gauss, J. *Mol. Phys.* **2008**, *106*, 1181–1192.
- (19) Nichols, P.; Govind, N.; Bylaska, E. J.; de Jong, W. A. *J. Chem. Theory Comput.* **2009**, *5*, 491–499.
- (20) Iikura, H.; Tsuneda, T.; Yanai, T.; Hirao, K. *J. Chem. Phys.* **2001**, *115*, 3540–3544.
- (21) Yanai, T.; Tew, D. P.; Handy, N. C. *Chem. Phys. Lett.* **2004**, *393*, 51–57.
- (22) Livshits, E.; Baer, R. *Phys. Chem. Chem. Phys.* **2007**, *9*, 2932–2941.
- (23) Govind, N.; Valiev, M.; Jensen, L.; Kowalski, K. *J. Phys. Chem. A* **2009**, *113*, 6041.
- (24) Jensen, L.; Govind, N. *J. Phys. Chem. A* **2009**, *113*, 9761.
- (25) Bast, R.; Schwerdtfeger, P. *J. Chem. Phys.* **2003**, *119*, 5988–5994.
- (26) van Lenthe, E.; van Lingen, J. N. *Int. J. Quantum Chem.* **2006**, *106*, 2525–2528.
- (27) van Lenthe, E.; van der Avoird, A.; Hagen, W. R.; Reiijerse, E. J. *J. Phys. Chem. A* **2000**, *104*, 2070–2077.
- (28) Faas, S.; van Lenthe, J. H.; Hennum, A. C.; Snijders, J. G. *J. Chem. Phys.* **2000**, *113*, 4052–4059.
- (29) Schwarz, W. H. E. Fundamentals of Relativistic Effects in Chemistry. In *The Concept of the Chemical Bond*; Masic, Z. B., Ed.; Springer: Berlin, 1990; Vol. 2, pp 559–643.
- (30) Kellö, V.; Sadlej, A. J. *Int. J. Quantum Chem.* **1998**, *68*, 159–174.
- (31) Malkin, I.; Malkina, O. L.; Malin, V. G. *Chem. Phys. Lett.* **2002**, *361*, 231–236.
- (32) Baerends, E. J.; Schwerdtfeger, P.; Snijders, J. G. *J. Phys. B* **1990**, *23*, 3225–3240.
- (33) van Lenthe, E. Ph.D. thesis, Vrije Universiteit Amsterdam, Netherlands, 1996.
- (34) Faas, S. Ph.D. thesis, Rijksuniversiteit Groningen, Netherlands, 2000.
- (35) Boys, S. F. *Proc. R. Soc. London* **1950**, *200*, 542–554.
- (36) King, H. F.; Dupuis, M. *J. Comput. Phys.* **1976**, *21*, 144–165.
- (37) Matsuoka, O. *Int. J. Quantum Chem.* **1971**, *5*, 1–11.
- (38) Taketa, H.; Huzinaga, S.; O-ohata, K. *J. Phys. Soc. Jpn.* **1966**, *21*, 2313–2324.
- (39) Helgaker, T.; Watson, M.; Handy, N. C. *J. Chem. Phys.* **2000**, *113*, 9402–9409.
- (40) Dupuis, M. *Comput. Phys. Commun.* **2001**, *134*, 150–166.
- (41) Obara, S.; Saika, A. *J. Chem. Phys.* **1986**, *84*, 3963.
- (42) Bylaska, E. J. *NWChem*, version 5.1; Pacific Northwest National Laboratory: Richland, WA, 2007.
- (43) Kendall, R. A.; Apra, E.; Bernholdt, D. E.; Bylaska, E. J.; Dupuis, M.; Fann, G. I.; Harrison, R. J.; Ju, J.; Nichols, J. A.; Nieplocha, J.; Straatsma, T. P.; Windus, T. L.; Wong, A. T. *Comput. Phys. Commun.* **2000**, *128*, 260–283.
- (44) Valiev, M.; Bylaska, E. J.; Govind, N.; Kowalski, K.; Straatsma, T. P.; van Dam, H. J. J.; Wang, D.; Nieplocha, J.; Apra, E.; Windus, T. L.; de Jong, W. A. *Comput. Phys. Commun.* **2010**, *181*, 1477–1489.
- (45) Weigend, F.; Ahlrichs, R. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3295–3305.
- (46) Feller, D. *J. Comput. Chem.* **2000**, *17*, 1571–1586.

- (47) Schuchardt, K.; Didier, B.; Elsethagen, T.; Sun, L.; Gurmooorthi, V.; Chase, J.; Li, J.; Windus, T. *J. Chem. Inf.* **2007**, *47*, 1045–1052.
- (48) Belanzoni, P.; Baerends, E.; van Lenthe, E. *Mol. Phys.* **2005**, *103*, 775–787.
- (49) Tsuchiya, T.; Abe, M.; Nakajima, T.; Hirao, K. *J. Chem. Phys.* **2001**, *115*, 4463–4472.
- (50) Pantazis, D. A.; Neese, F. *J. Chem. Theory Comput.* **2009**, *5*, 2229–2238.
- (51) Roos, B. O.; Lindh, R.; Malmqvist, P.; Veryazov, V.; Widmark, P. *J. Phys. Chem. A* **2005**, *109*, 6575–6579.
- (52) Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098–3100.
- (53) Perdew, J. P. *Phys. Rev. B* **1986**, *33*, 8822–8824.
- (54) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785–789.
- (55) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 1372–1377.
- (56) Thierfelder, C.; Schwerdtfeger, P.; Saue, T. *Phys. Rev. A* **2007**, *76*, 034502–4.
- (57) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865–3868.
- (58) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1997**, *78*, 1396.
- (59) Zhang, Y.; Yang, W. *Phys. Rev. Lett.* **1998**, *80*, 890.
- (60) Glendening, E. D.; Badenhop, J. K.; Reed, A. E.; Carpenter, J. E.; Bohmann, J. A.; Morales, C. M.; Weinhold, F. *NBO 5.0*; Theoretical Chemistry Institute, University of Wisconsin: Madison, WI, 2001. <http://www.chem.wisc.edu/~nbo> 5 (accessed July 2010).
- (61) Ye, A.; Autschbach, J. *J. Chem. Phys.* **2006**, *125*, 234101–13.
- (62) Autschbach, J.; Zheng, S. *Magn. Reson. Chem.* **2008**, *46*, S48–S55.
- (63) van Wullen, C. *J. Chem. Phys.* **1998**, *109*, 392–399.
- (64) Philipsen, P. H. T.; van Lenthe, E.; Snijders, J. G.; Baerends, E. *J. Phys. Rev. B* **1997**, *56*, 13556–13562.
- (65) Hensel, K. D.; Styger, C.; Jager, W.; Merer, A. J.; Gerry, M. C. L. *J. Chem. Phys.* **1993**, *99*, 3320–3328.
- (66) Lucken, E. A. C. In *Advances in Nuclear Quadrupole Resonance*; Smith, J. A. S., Ed.; Wiley: New York, 1983; Vol. 5, pp 83–124.
- (67) Palmer, M. H. *Naturforsch. A* **1998**, *53*, 615.
- (68) Walker, K. A.; Gerry, M. C. *J. Mol. Spectrosc.* **1999**, *193*, 224–227.
- (69) Sheridan, J. In *Advances in Nuclear Quadrupole Resonance*; Smith, J. A. S., Ed.; Wiley: New York, 1983; Vol. 5, pp 125–163.
- (70) Gordy, W.; Cook, R. L. *Microwave Molecular Spectra*; Wiley: New York, 1984; pp 859–872.
- (71) Pyykkö, P. *Mol. Phys.* **2008**, *106*, 1965–1974.
- (72) Schwerdtfeger, P.; Pernpointner, M.; Nazarewicz, W. Calculation of nuclear quadrupole coupling constants. In *Calculation of NMR and EPR parameters: Theory and applications*; Kaupp, M., Bühl, M., Malkin, V. G., Eds.; Wiley-VCH: Weinheim, Germany, 2004; pp 279–291.
- (73) Schwerdtfeger, P.; Bast, R.; Gerry, M. C. L.; Jacob, C. R.; Jansen, M.; Kellö, V.; Mudring, A. V.; Sadlej, A. J.; Saue, T.; Söhnel, T.; Wagner, F. E. *J. Chem. Phys.* **2005**, *122*, 124317.
- (74) Schwerdtfeger, P.; Pernpointner, M.; Laerdahl, J. K. *J. Chem. Phys.* **1999**, *111*, 3357–3364.
- (75) Townes, C. H.; Dailey, B. P. *J. Chem. Phys.* **1949**, *17*, 782–796.
- (76) Szilagy, R. K.; Metz, M.; Solomon, E. *J. Phys. Chem. A* **2002**, *106*, 2994–3007.
- (77) Sanderson, C. T.; Palmer, B. J.; Morgan, A.; Murphy, M.; Dluhy, R. A.; Mize, T.; Amster, I. J.; Kutal, C. *Macromolecules* **2002**, *35*, 9648–9652.
- (78) Bard, A. J.; Fox, M. A. *Acc. Chem. Res.* **1995**, *28*, 141–145.
- (79) Hara, M.; Waraksa, C. C.; Lean, J. T.; Lewis, B. A.; Mallouk, T. E. *J. Phys. Chem. A* **2000**, *104*, 5275–5280.
- (80) Gao, F. G.; Bard, A. J. *J. Am. Chem. Soc.* **2000**, *122*, 7426–7427.
- (81) Hagfeldt, A.; Graetzel, M. *Chem. Rev.* **1995**, *95*, 49–68.
- (82) Klimant, I.; Wolfbeis, O. S. *Anal. Chem.* **1995**, *67*, 3160–3166.
- (83) Erkkila, K. E.; Odom, D. T.; Barton, J. K. *Chem. Rev.* **1999**, *99*, 2777–2795.
- (84) Armistead, P. M.; Thorp, H. H. *Bioconjugate Chem.* **2002**, *13*, 172–176.
- (85) Kirsch-De Mesmaeker, A.; Lecomte, J. P.; Kelly, J. M. *Top. Curr. Chem.* **1996**, *177*, 25–76.
- (86) Barton, J. K. *Science* **1986**, *233*, 727–734.
- (87) Sundquist, W. I.; Lippard, S. J. *Coord. Chem. Rev.* **1990**, *100*, 293–322.
- (88) Pauly, M.; Kayser, I.; Schmitz, M.; Dicato, M.; Del Guerso, A.; Kolber, I.; Moucheron, C.; Kirsch-De Mesmaeker, A. *Chem. Commun.* **2002**, *108*, 6–1087.
- (89) Pourtois, G.; Belionne, D.; Moucheron, C.; Schumm, S.; Kirsch-De Mesmaeker, A.; Lazzaroni, R.; Bredas, J.-L. *J. Am. Chem. Soc.* **2004**, *126*, 683–692.
- (90) Bechthold, H.; Rehder, D. *J. Organomet. Chem.* **1981**, *206*, 305–315.
- (91) Bühl, M.; Gaemers, S.; Elsevier, C. J. *Chem.—Eur. J.* **2000**, *6*, 3272–3280.
- (92) Wagner, F. E.; Wagner, U. *Mössbauer Isomer Shifts*, 1st ed.; Shenoy, G. K., Wagner, F. E., Eds.; North Holland: Amsterdam, 1978; pp 431–514.
- (93) Lo, A. Y. H.; Bitterwolf, T. E.; Macdonald, C. L. B.; Schurko, R. W. *J. Phys. Chem. A* **2005**, *109*, 7073–7087.
- (94) Schreckenbach, G.; Shamov, G. A. *Acc. Chem. Res.* **2010**, *43*, 19–29.
- (95) Cho, H.; de Jong, W. A.; Soderquist, C. Z. *J. Chem. Phys.* **2010**, *132*, 084501.
- (96) Pepper, M.; Bursten, B. E. *Chem. Rev. B* **1991**, *91*, 719–741.
- (97) Clark, D. L.; Hobart, D. E.; Neu, M. P. *Chem. Rev.* **1995**, *95*, 25–48.
- (98) de Jong, W. A.; Visscher, L.; Nieuwpoort, W. C. *J. Mol. Struct.* **1998**, *458*, 41–52.
- (99) Monard, J. A.; Huray, P. G.; Thomson, J. O. *Phys. Rev. B* **1974**, *9*, 2838–2845.
- (100) Larsson, S.; Pyykkö, P. *Chem. Phys.* **1986**, *101*, 355.
- (101) Pyykkö, P.; Seth, M. *Theor. Chem. Acc.* **1997**, *96*, 92.

π Interactions Studied with Electronic Structure Methods: The Ethyne Methyl Isocyanide Complex and Thioanisole

Natalie H. Bretherick and Tanja van Mourik*

School of Chemistry, University of St. Andrews, North Haugh, St. Andrews, Fife, KY16 9ST, Scotland, United Kingdom

Received June 03, 2010

Abstract: Two molecular systems for which previous studies had found qualitative differences in the results from calculations with the B3LYP and MP2 methods are investigated with a range of different electronic structure methods, including meta and double hybrid density functionals and DFT-D (DFT augmented with an empirical dispersion term). The performance of the different methods is assessed by comparison to estimated CCSD(T)/CBS (complete basis set) results. The first molecular system studied is the ethyne methyl isocyanide complex ($\text{CH}_3\text{NC}-\text{C}_2\text{H}_2$), which exhibits π hydrogen bonds involving the $\text{C}\equiv\text{C}$ and $\text{N}\equiv\text{C}$ triple bonds. Earlier work on this system had shown that B3LYP predicts significantly longer hydrogen-bond distances than MP2. Here, we show that this is likely due to missing dispersion in the B3LYP calculations. On the basis of the CCSD(T) results, the ethyne methyl isocyanide interaction energy is estimated to be 12 ± 1 kJ/mol. B3LYP significantly underestimates the stability of the complex, whereas MP2 slightly overestimates. M05-2X, B3LYP-D, and (CP-corrected) mPW2-PLYP-D give results in close proximity to the CCSD(T) reference values. The second molecule investigated is thioanisole ($\text{C}_6\text{H}_6\text{SCH}_3$), which can adopt two different conformations (thiomethyl group either planar or perpendicular with respect to the benzene ring). Potential energy curves for rotation around the $\text{C}(\text{sp}^2)-\text{S}$ bond are computed and compared to the estimated CCSD(T)/CBS curve. CCSD(T) predicts the planar conformation to be the global minimum, with a plateau region near the perpendicular conformation (~ 4 kJ/mol higher in energy than the planar conformation). The shape of the curve, and location of minima and barriers, is very dependent on the method and basis set employed. MP2, B3LYP, M05-2X, mPW2-PLYP, and mPW2-PLYP-D (employing basis sets of double- or triple- ζ quality) give results in reasonable agreement with the CCSD(T) results, whereas B3LYP-D and M06-L give vastly overestimated barriers at the perpendicular conformation.

1. Introduction

Interactions with π systems are important in many areas of chemistry. For example, interactions with the ring systems of the aromatic amino acids tyrosine, phenylalanine, and tryptophan stabilize peptides and proteins,¹ whereas π -stacking interactions between consecutive nucleic acid bases stabilize the structure of the DNA double helix.² Interactions with aromatic rings also play key roles in chemical and biological recognition³ and supramolecular chemistry.^{4,5} A

computational description of these interactions is however fraught with difficulties, because π interactions are much weaker than covalent interactions. In addition, London dispersion forces play an important role in π interactions, and these are inherently difficult to describe properly using computational methods. Density functional theory (DFT) has become very popular over the past few decades, due to its greater computational efficiency relative to correlated ab initio methods like second-order Møller–Plesset (MP2) perturbation theory.⁶ However, conventional density functionals like B3LYP^{7–9} do not describe dispersion^{10–22} appropriately and may therefore not give correct results for

* Corresponding author e-mail: tanja.vanmourik@st-andrews.ac.uk.

π interactions. Examples of the deficiency of the B3LYP functional to correctly describe π interactions include the π -bonded indole–water minimum,²³ and some conformers of the tyrosine–glycine (Tyr–Gly) dipeptide.²⁴ MP2 does describe dispersion; however, MP2 calculations suffer from larger basis set superposition error (BSSE) values than DFT, particularly when small to medium-sized basis sets are employed, and may therefore also not produce the correct result. For example, we found before that BSSE distorts the MP2/6-31+G(d) potential energy surface of some Tyr–Gly conformers to such an extent that artificial minima are created or real minima are masked.^{24,25}

The current study evolved from two recent computational studies in the literature,^{26,27} reporting qualitative differences in the results obtained with the MP2 and B3LYP methods. For both molecular systems investigated in these studies, π interactions are likely important. One of these is an intermolecular complex, ethyne methyl isocyanide ($\text{CH}_3\text{NC}-\text{C}_2\text{H}_2$), which exhibits π hydrogen bonds (H bonds) involving the $\text{C}\equiv\text{C}$ and $\text{N}\equiv\text{C}$ triple bonds. For this system, MP2 predicts a much closer contact between the two molecules in the complex and yields a stronger interaction than B3LYP.²⁶ The second molecular system is thioanisole, consisting of a benzene ring and a thiomethyl ($-\text{SCH}_3$) moiety. This molecule may exhibit intramolecular interactions between the thiomethyl group and the π -electron cloud of the benzene ring. Here, B3LYP and MP2 were reported to give different minima.²⁷ B3LYP/cc-pVTZ yields a planar structure, in which the thiomethyl group and the benzene ring are coplanar, whereas MP2/6-311G(d,p) calculations result in a conformer in which the thiomethyl group is perpendicular with respect to the aromatic ring. MP2/cc-pVTZ finds both structures, with the planar one the most stable of these.

There have been a number of previous computational studies on the structure of thioanisole. Vondrák et al.²⁸ computed a potential energy curve for rotation around the $\text{C}(\text{sp}^2)\text{-S}$ bond using Hartree–Fock (HF) and the 6-311G(d,p) basis set and found a single potential energy minimum at the perpendicular conformation. Dal Colle et al.²⁹ found similar results using HF/6-31G(d,p). Dolgounitcheva et al.³⁰ optimized the structure of the molecule at the HF and B3LYP levels of theory, using the 6-31G(d) and (for HF) 6-311G(d) basis sets. HF optimizations yielded minima for the perpendicular conformation and transition states for the planar structure. B3LYP optimizations yielded the opposite result. Gellini et al.³¹ also located the planar conformer with B3LYP/6-31G(d,p). Bzhezovskii and Kapustin^{32,33} computed the potential energy curves using B3LYP/6-31G(d) and MP2/6-31G(d). The B3LYP calculations found two minima at the planar and perpendicular conformations, with the planar one the global minimum. In contrast, MP2 located the minimum near the planar conformation, whereas both the planar and perpendicular structures were characterized as transition states. Bossa et al.³⁴ computed potential energy curves using HF, MP n ($n = 2-4$), and various density functionals, including B1LYP, B3LYP, mPW1PW, and Bh&hLYP. The MP n calculations employing a 6-31G(d) basis set found that the perpendicular structure is a minimum whereas the planar

structure is a transition state. The DFT calculations gave different results, yielding the planar structure as the global minimum. Some functionals located a secondary minimum close to the perpendicular conformation. The perpendicular conformation was found to be a transition state for all functionals employed. Apart from the study of Suzuki et al.,²⁷ which prompted the current work, we are aware of only one study that employed basis sets larger than 6-31G(d) in the MP2 calculations: Shiskov et al.³⁵ found a single minimum for the perpendicular conformation with MP2/6-31G(d), an additional shallow minimum for the planar conformation with MP2/6-311G(d,p), whereas MP2/cc-pVTZ calculations yield the planar orientation as the global minimum, with an additional shallow minimum at the perpendicular orientation. In agreement with the previous studies, B3LYP/cc-pVTZ predicts a single minimum for the planar orientation.

Numerous experimental studies on thioanisole^{28,29,31,35-50} indicate the existence of a single planar or nearly planar structure, or the coexistence of the planar and perpendicular forms. B3LYP calculations indicate that the barrier height for rotation around the $\text{C}(\text{sp}^2)\text{-S}$ bond is much smaller for thioanisole than for anisole.³¹ Solution-phase NMR predicts a barrier height for the thioanisole internal rotation of 6.0 kJ/mol,^{44,45,47} whereas various computational studies showed that the barrier (both the height as well as the location) is very dependent on the computational method employed.^{27,33-35}

Only a few studies on molecules with the type of π H bonds as occurring in the ethyne methyl isocyanide complex ($\text{C}\equiv\text{N}$ bond acting as the H-bond acceptor) have been published in the literature. Using the B3LYP and MP2 methods, Bakó et al.⁵¹ found two distinct minima for the acetonitrile–water complex; one of these contains a H bond between one of the water hydrogens and the $\text{C}\equiv\text{N}$ bond of acetonitrile. The MP2 calculations yielded $\text{H}\cdots\text{C}$ and $\text{H}\cdots\text{N}$ distances that are about 0.1 Å shorter than those predicted by B3LYP. Cao et al.²⁶ studied several of these systems, including $\text{CH}_3\text{C}\equiv\text{N}\cdots\text{H}_2\text{O}$, $\text{CH}_3\text{N}\equiv\text{C}\cdots\text{H}_2\text{O}$, $\text{CH}_3\text{C}\equiv\text{N}\cdots\text{H}_3\text{N}$, and $\text{CH}_3\text{N}\equiv\text{C}\cdots\text{H}_2\text{C}_2$ (ethyne methyl isocyanide). For all of these, MP2 predicts shorter intermolecular distances than B3LYP. The differences are largest in the ethyne methyl isocyanide complex, where the $(\text{C}=\text{C})\cdots\text{H}(\text{CH}_2)$ distance predicted by MP2 is nearly 0.5 Å shorter than the value obtained with B3LYP. Tschumper et al.⁵²⁻⁵⁴ studied a number of $\pi\text{-}\pi$ systems, including the dimers and trimers of cyanogen ($\text{N}\equiv\text{C}-\text{C}\equiv\text{N}$). They showed that an accurate description of the π interactions in these systems requires the application of the CCSD(T) method (or the equivalent thereof). In addition, even the inclusion of quadruple excitations was found to result in non-negligible changes to the binding energy.⁵³

In the current paper, calculations using CCSD(T) (coupled cluster including single, double, and perturbative triple excitations), at the estimated CBS (complete basis set) limit, are used to assess the correctness of the B3LYP and MP2 methods to describe the two molecular systems studied in this work. CCSD(T) has been described as the “gold standard” of electronic structure theory, providing very reliable results for basically all single-reference systems. In addition, several other electronic structure methods are

assessed, including the meta and meta hybrid functionals M06-L⁵⁵ and M05-2X,⁵⁶ the double hybrid functional mPW2-PLYP,^{57,58} and DFT-D (DFT augmented with an empirical dispersion term).^{59,60} M06-L and M05-2X were developed to provide broad applicability in chemistry and have been shown to yield much improved results for weak intermolecular interactions as compared to B3LYP.^{55,61–74} Also, DFT-D generally yields good results for weak interactions,^{64,74–78} though it has been shown that B3LYP-D yields overestimated interaction energies for anisole–water and anisole–ammonia.⁷⁹ The double hybrid functional mPW2-PLYP was found to give good performance in general,^{57,80} though it underestimates the interaction in van der Waals complexes.^{81,82} mPW2-PLYP satisfactorily describes the well depth of CO₂–He and CO₂–Ne but fails to produce correct interaction energies for CO₂ complexed with heavier rare-gas atoms.⁸³ To overcome the underestimated dispersion in mPW2-PLYP, we also consider the mPW2-PLYP-D method (mPW2-PLYP augmented with an empirical dispersion term).⁸¹ This method was found to give the best performance for the description of the three minima along the ϕ_{Gly} rotational profile of one particular Tyr-Gly conformer.⁶⁴

As mentioned above, BSSE may be a possible explanation for the different results obtained with the B3LYP and MP2 methods. It is well recognized that BSSE causes an overestimation of the stability of intermolecular complexes, like the ethyne methyl isocyanide complex. However, intramolecular BSSE can also have effects on the relative conformational energies of single molecules. In thioanisole, close contact between the benzene ring and the –SCH₃ group may induce intramolecular BSSE. For both molecular systems, the effects of BSSE are investigated using the counterpoise (CP) procedure of Boys and Bernardi.⁸⁴ The treatment of the *intermolecular* BSSE in the ethyne methyl isocyanide complex is straightforward: the standard procedure is to take the individual monomers, in this case the ethyne and methyl isocyanide molecules, as the isolated fragments. However, correction of the *intramolecular* BSSE in the thioanisole molecule is more ambiguous. In previous work,^{24,25,85} the intramolecular BSSE in structures of the tyrosyl–glycine (Tyr-Gly) dipeptide was approximated by the intermolecular BSSE in complexes of phenol and *N*-formylglycine with conformations and spatial arrangements identical to the Tyr-Gly structures. Thus, the Tyr-Gly molecule was split into two fragments by removing the –CH₂–N(H)(NH₂)– linkage between the phenol and *N*-formylglycine moieties, after which the dangling bonds were saturated with hydrogens. A similar approach was used by Valdés et al.⁸⁶ to compute the intramolecular BSSE in different conformations of the phenylalanine–glycine–phenylalanine tripeptide, which was partitioned into two benzene molecules with the same geometries and spatial arrangements as in the tripeptide. In thioanisole, we want to estimate the magnitude of the intramolecular BSSE caused by interaction between the benzene ring and the –SCH₃ group. However, there is no connecting group between these two fragments, and the same CP procedure as used for Tyr-Gly is therefore not feasible here. In the current work, we simply break the C(sp²)–S

bond, creating two open-shell radical molecules to be used as the monomers in the CP procedure. This procedure is similar to the intramolecular counterpoise correction employed previously for the N₂⁸⁷ and HCl⁸⁸ molecules.

Other schemes to correct intramolecular BSSE have been proposed. Palermo et al. published the so-called rotation method.⁸⁹ According to this method, for a molecule with two interacting groups that may be affected by intramolecular BSSE, one of the interacting groups is rotated to a position where no interaction is observed, and ghost orbitals are placed corresponding to the rotated group in its original position. In addition, a geometry is created by fixing the interacting group in its original, optimal position and placing ghost orbitals corresponding to the group in its rotated position. The interaction energy is then obtained by taking the difference in energy of these two geometries. It was however shown that this method underestimates the magnitude of the intramolecular BSSE.⁹⁰ In addition, this method cannot be easily used to compute BSSE-free rotational energy profiles. Other proposed schemes to correct intramolecular BSSE are all variants of the CP procedure. Jensen⁹¹ investigated intramolecular BSSE in different conformations of H₂O, NH₃, and ethane by taking the union of the basis sets of the two different conformations. However, this scheme cannot be used when the atoms in the two conformations are in similar positions, as the ghost orbitals would be very close to atomic basis functions, leading to numerical problems. More recently, the same author devised an atomic counterpoise scheme where the BSSE is estimated as a sum of atomic contributions, calculated as differences in energies computed in a regular basis set and in a subset of basis functions on atoms separated by a minimum number of bonds.⁹² This method does not require the somewhat ambiguous definition of the molecular fragments as in the CP approach used for thioanisole in the current paper. However, the atomic counterpoise method does require the definition of the atomic reference states and the subset of basis functions to be included in the CP calculations and is therefore also not explicitly defined. Baladin⁹³ used the close relationship between intramolecular BSSE and basis set incompleteness to derive a power-law linking basis set incompleteness and intramolecular BSSE. This could in principle be used to obtain MP2/CBS values using only one energy value and the intramolecular BSSE (or, alternatively, compute the intramolecular BSSE if the CBS limit is known). However, the law has a large uncertainty (25%) and has so far only been tested on a number of small molecules (ranging from N₂ to C₂H₄) using nondiffuse basis sets. Salvador et al.^{94,95} showed that intramolecular BSSE is responsible for the surprising result that *ab initio* calculations on aromatic systems such as benzene and the nucleic acid bases yield nonplanar structures. They corrected the intramolecular BSSE in these molecules using small moieties (the C–H, C=O, N–H, and CCH₃ groups) as the counterpoise fragments (while noting that the unambiguous, but expensive, way to do the counterpoise calculations would be to take atomic fragments). A similar method, employing CH₂ and CH₃ fragments, was used to estimate the intramolecular BSSE in different conformations of *n*-butane and *n*-hexane.⁹⁶ This

approach is basically identical to the one adopted in the current paper to compute the intramolecular BSSE in thioanisole.

In the current paper, we show that the discrepancies between the B3LYP and MP2 results for the ethyne methyl isocyanide complex are due primarily to missing dispersion in the B3LYP calculations and, to a smaller extent, overestimation of the interaction by the MP2 method. This molecule shows the typical characteristics of a dispersion-bound system: the binding energy is underestimated by B3LYP and slightly overestimated by MP2, and good performance is achieved with the M05-2X, DFT-D, and mPW2-PLYP-D methods. The situation is very different for thioanisole. We find that the shape of the potential energy curve for rotation around the thioanisole C(sp²)-S bond is very method- and basis-set-dependent. The CCSD(T) reference profile shows the global minimum at the planar conformation and a plateau region at the perpendicular conformation. This can be reasonably reproduced by MP2 calculations with basis sets of at least triple- ζ quality and B3LYP calculations with basis sets of double- ζ quality. The electronic structure methods assessed show a very mixed performance: B3LYP-D and M06-L considerably overestimate the relative stability of the planar conformer, whereas the M05-2X and mPW2-PLYP(-D) results are in closer agreement with the CCSD(T) reference values. This shows that DFT methods designed to give good results for weak interactions, like DFT-D and the M0x ($x = 5, 6$) functionals, do not always give good results, and more studies assessing the performance of these methods are required, particularly for challenging molecular systems like thioanisole.

2. Methodology

2.1. The Ethyne Methyl Isocyanide Complex. The ethyne methyl isocyanide complex was optimized using B3LYP, MP2, and M05-2X, all employing the aug-cc-pVTZ basis set. The geometry optimizations were done both with and without counterpoise (CP) corrections, and the interaction energy was computed according to the CP procedure.⁸⁴ The CP-corrected interaction energy follows from:

$$\Delta E^{\text{CP}} = E_{\text{EM}}^{\{\text{EM}\}} - E_{\text{E}}^{\{\text{E}\}} - E_{\text{M}}^{\{\text{M}\}} + E_{\text{E}}^{\text{def}} + E_{\text{M}}^{\text{def}} \quad (1)$$

Here, the monomer deformation energy $E_{\text{X}}^{\text{def}}$ ($\text{X} = \text{E}$ or M ; $\text{E} = \text{ethyne}$; $\text{M} = \text{methyl isocyanide}$) is defined as

$$E_{\text{X}}^{\text{def}} = E_{\text{X}}^{\{\text{X}\}}(\text{EM}) - E_{\text{X}}^{\{\text{X}\}}(\text{X}) \quad (2)$$

In eqs 1 and 2, the subscripts denote the molecular system. The superscripts denote the basis set used for the calculation (the monomer basis set $\{\text{E}\}$ or $\{\text{M}\}$ or the dimer basis set $\{\text{EM}\}$), whereas the terms in parentheses indicate the geometry used in the calculation (the geometry of the dimer (EM) or the optimized monomer geometry (E) or (M)).

Interaction energy profiles were computed by optimizing the structures at fixed C₃-N₆ distances in the range 2.0–4.5 Å, using B3LYP/aug-cc-pVTZ, MP2/6-31+G(d), and MP2/aug-cc-pVTZ (see Figure 1 for the atom labeling). The

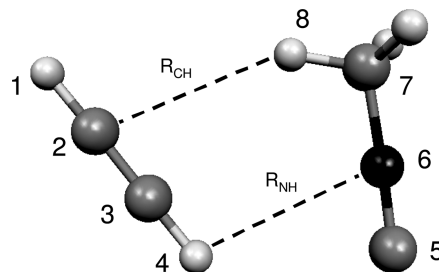


Figure 1. Atom labeling in the ethyne methyl isocyanide complex.

geometry optimizations were performed on the uncorrected potential energy surfaces. Uncorrected and CP-corrected interaction energies were computed at the same levels of theory as used to obtain the optimized structures.

The uncorrected B3LYP/aug-cc-pVTZ-optimized geometry was used to compare the performance of a range of electronic structure methods for evaluating the ethyne methyl isocyanide interaction energy. Uncorrected as well as CP-corrected interaction energies were computed. To avoid artificially large monomer deformation energies, the calculation of the monomer energies $E_{\text{E}}^{\{\text{E}\}}(\text{E})$ and $E_{\text{M}}^{\{\text{M}\}}(\text{M})$ used the B3LYP/aug-cc-pVTZ-optimized geometries. The levels of theory considered include HF, B3LYP, MP2, M05-2X, M06-L, B3LYP-D, and mPW2-PLYP-D, all employing the aug-cc-pVTZ basis set. The MP2 calculations were also done with the aug-cc-pVDZ and aug-cc-pVQZ basis sets. From the mPW2-PLYP-D calculations, the mPW-LYP and mPW2-PLYP interaction energies can be extracted as well. mPW-LYP is a nonhybrid functional.

The results were compared with the estimated CCSD(T)/CBS interaction energy, which was obtained by subtracting the CCSD(T)/CBS monomer energies from the CCSD(T)/CBS dimer energy. As above, the dimer and monomer structures were optimized with B3LYP/aug-cc-pVTZ. The interaction energy obtained like this is free of BSSE, as the BSSE vanishes at the CBS limit. The CCSD(T)/CBS energies were obtained by the method introduced by Klopper and Lüthi⁹⁷ and used by Hobza and co-workers,⁹⁸ which involves the addition of a higher-order correlation correction term to the MP2/CBS energies, obtained by extrapolation. The CCSD(T)/CBS dimer and monomer energies then follow from

$$E(\text{CCSD(T)/CBS}) = E(\text{HF/av5z}) + E(\text{MP2corr/CBS(avtz/avqz)}) + E(\text{CCSD(T)corr/avtz}) \quad (3)$$

In eq 3, the aug-cc-pVxZ basis sets are abbreviated as avxz ($x = \text{t}, \text{q}, 5$). The MP2 correlation energy term, $\Delta E(\text{MP2corr/CBS})$, was obtained by extrapolating the aug-cc-pVTZ and aug-cc-pVQZ MP2 correlation energies to the CBS limit using the extrapolation formula of Halkier et al.⁹⁹

$$E(\text{MP2corr/CBS}) = \frac{x^3}{x^3 - (x-1)^3} \times E_{\text{MP2corr},x} - \frac{(x-1)^3}{x^3 - (x-1)^3} E_{\text{MP2corr},x-1} \quad (4)$$

Table 1. Interaction Energies ΔE (in kJ/mol) and Key Structural Features (Distances in Å, Angles in Degrees) of the Ethyne Methyl Isocyanide Complex^a

	geometry ^b	ΔE^{noCP}	ΔE^{CP}	$R(\text{H}_4\text{C}_5)$	$\angle(\text{H}_4\text{C}_5\text{N}_6)$	$R(\text{H}_8\text{C}_2)$	$\angle(\text{C}_7\text{H}_8\text{C}_2)$	$R(\text{C}_3\text{N}_6)$
B3LYP	no CP	-5.25	-5.07	2.941	84.2	3.407	147.0	3.602
	CP	-5.24	-5.07	2.959	84.2	3.400	147.3	3.611
MP2	no CP	-14.32	-12.62	2.759	83.8	2.916	150.2	3.305
	CP	-14.26	-12.68	2.790	84.9	2.972	149.1	3.341
M05-2X	no CP	-11.12	-10.80	2.836	83.3	3.057	149.8	3.402
	CP	-11.12	-10.80	2.838	83.3	3.066	149.6	3.407

^a All calculations employed the aug-cc-pVTZ basis set. ^b No CP/CP: geometries optimized on the uncorrected/CP-corrected potential energy surface.

Here, x is the cardinal number of the largest basis set used in the extrapolation (in this case, 4 for aug-cc-pVQZ).

The CCSD(T) correction term for higher-order correlation energy, $E(\text{CCSD(T)corr})$, was obtained by taking the difference in the MP2 and CCSD(T) correlation energies, computed with the aug-cc-pVTZ basis set. As the basis set dependence of the CCSD(T) correction term is small,^{100–102} the aug-cc-pVTZ basis set should be sufficiently large to give an accurate estimate of the higher-order correlation energy.

The B3LYP, MP2, and M05-2X calculations were done with Gaussian,¹⁰³ the mPW2-PLYP-D and B3LYP-D calculations were done with Orca,¹⁰⁴ and the Molpro program package¹⁰⁵ was used for the CCSD(T) calculations. For consistency with the Gaussian results, the B3LYP-D calculations with Orca employed the VWN1¹⁰⁶ correlation functional (Gaussian's definition of the B3LYP functional).

2.2. Thioanisole. Potential energy curves for rotation around the $\text{C}(\text{sp}^2)\text{--S}$ bond were computed by optimizing the thioanisole structure at fixed $\tau(\text{CCSC})$ torsion angles ranging from 0 (planar) to 90° (perpendicular) using M05-2X/6-31+G(d). Single-point calculations employing the M05-2X-optimized geometries were performed at different levels of theory, including MP2 and B3LYP employing the 6-311G(d,p) and (aug-)cc-pV x Z ($x = \text{D, T, Q}$) basis sets; HF/cc-pV x Z ($x = \text{D, T, Q, 5}$); CCSD(T)/aug-cc-pVDZ; B3LYP-D/aug-cc-pVDZ; and the M05-2X, M06-L, and mPW2-PLYP(-D) methods combined with the (aug-)cc-pV x Z ($x = \text{D, T}$) basis sets. M05-2X and M06-L calculations were also done with the aug-cc-pV($n+d$)Z ($n = \text{D, T}$)¹⁰⁷ basis sets, obtained from the EMSL Basis Set Exchange Web site.^{108,109} An estimated CCSD(T)/CBS potential energy curve was constructed by estimating the CCSD(T)/CBS thioanisole energy at each $\tau(\text{CCSC})$ value similar to eq 3, but at slightly lower basis set levels:

$$E(\text{CCSD(T)/CBS}) = E(\text{HF/v5z}) + E(\text{MP2corr/CBS(avtz/avqz)}) + E(\text{CCSD(T)corr/avdz}) \quad (5)$$

Selected single-point energies were CP-corrected by considering as the monomers the two radical fragments obtained by breaking the $\text{C}(\text{sp}^2)\text{--S}$ bond, i.e., the benzene ring and the --SCH_3 group.

3. Results

3.1. The Ethyne Methyl Isocyanide Complex. *3.1.1. Geometry Optimization.* Table 1 shows the interaction energies and key structural features of the ethyne methyl isocyanide

complex optimized at different levels of theory. In agreement with the results of Cao et al.,²⁶ MP2 predicts shorter intermolecular distances than B3LYP. The small differences in the results obtained in ref 26 and by us may be due to different convergence criteria (we used Gaussian's "tight" convergence threshold) or different integration grid sizes in the DFT calculations (we used Gaussian's default integration grid). In line with the larger intermolecular distances, B3LYP predicts much less intermolecular stabilization than MP2. The M05-2X results are between the B3LYP and MP2 results but closer to the MP2 results.

The effect of CP correction on the geometries is very small. Even the MP2 geometries are only negligibly affected by BSSE: CP correction increases the intermolecular distances by less than 0.1 Å. As expected, the effects are even smaller for the DFT calculations. The interaction energies, on the other hand, are more noticeably affected by BSSE: CP correction decreases the MP2 interaction energy by about 10%. We conclude that CP-corrected geometry optimization is, in this case, not required, but the ethyne methyl isocyanide interaction energies should be corrected for BSSE.

These results suggest that the differences in the geometries optimized by B3LYP and MP2 are not due to large BSSE effects in the MP2 calculations. It is therefore likely that missing dispersion interactions in the B3LYP calculations are the cause of the discrepancies between the MP2 and B3LYP results. We explore this further in the next section.

3.1.2. Interaction Energy Profiles. To investigate further the effect of the method and BSSE on the intermolecular separation, we performed geometry optimizations at fixed $\text{C}_3\text{--N}_6$ distances, in the range 2.5–4.5 Å. At distances outside this range, the geometry did not resemble the minimum-energy structure displayed in Figure 1. Figure 2 shows the uncorrected and CP-corrected interaction energy profiles. The B3LYP/aug-cc-pVTZ profile is shown only up to 3.75 Å, as the ethyne methyl isocyanide structure changes significantly beyond this distance, yielding a structure slightly more stable than the minimum at 3.60 Å. Full optimization of this structure generated a linear complex, which was found to be more stable than the π -bonded structure investigated in this work. This conforms to the water–acetonitrile system, for which the linear H-bonded isomer was also found to be more stable than the π -bonded isomer.⁵¹

The inset of Figure 2 shows the variation of the BSSE with $\text{C}_3\text{--N}_6$ distance. The B3LYP/aug-cc-pVTZ BSSE is small for all $\text{C}_3\text{--N}_6$ distances considered, with the result that the uncorrected and CP-corrected B3LYP/aug-cc-pVTZ curves are nearly identical, and therefore, only the CP-

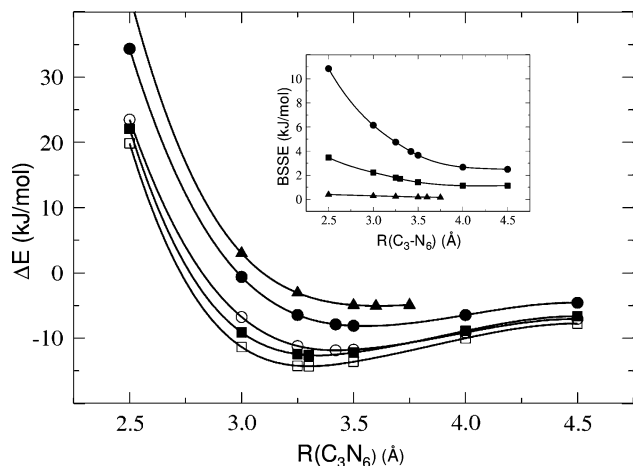


Figure 2. Interaction energy profiles obtained by optimizing the ethyne methyl isocyanide structure at fixed C_3-N_6 distances. Circles: MP2/6-31+G(d) profiles. Squares: MP2/aug-cc-pVTZ profiles. Triangles: B3LYP/aug-cc-pVTZ profile. Open symbols: uncorrected results. Closed symbols: CP-corrected results. The inset shows the variation of the BSSE as a function of the C_3-N_6 distance.

corrected profile is shown in Figure 2. The BSSE is somewhat larger in the MP2/aug-cc-pVTZ calculations and increases with decreasing separation between the two fragments. This is expected, as the basis-set stealing process, which underlies the cause of BSSE, becomes more difficult the further apart the fragments are. The MP2/aug-cc-pVTZ BSSE is however still sufficiently small not to affect significantly the interaction energy profiles; the CP-corrected MP2/aug-cc-pVTZ curve is just slightly shifted upward as compared to the uncorrected profile. These results are in agreement with the results in Table 1, which show a negligible effect of BSSE on the ethyne methyl isocyanide geometries obtained with DFT and MP2 employing the aug-cc-pVTZ basis set. However, the MP2/6-31+G(d) results in Figure 2 show much larger BSSE effects. The magnitude of the BSSE changes significantly over the C_3-N_6 distance range considered, and the application of the CP procedure shifts the minimum to a longer interfragment distance (uncorrected: $R(C_3N_6) = 3.423 \text{ \AA}$; CP-corrected: $R(C_3N_6) = 3.554 \text{ \AA}$). Thus, the effect of BSSE on the ethyne methyl isocyanide structure is only negligible if a sufficiently large basis set is employed in the MP2 calculations. In the current case, the aug-cc-pVTZ basis set fulfills this requirement.

3.1.3. Comparison of Different Electronic Structure Methods. Figure 3 shows a comparison of the ethyne methyl isocyanide interaction energy computed with different methods, using the uncorrected B3LYP/aug-cc-pVTZ optimized geometry. The estimated CCSD(T)/CBS interaction energy (-10.68 kJ/mol), obtained using eq 3, was used to assess the performance of the various methods.

As we used the estimated CCSD(T)/CBS interaction energy to assess other electronic structure methods, a knowledge of its accuracy is of considerable importance. The CCSD(T)/CBS interaction energy has three main sources of uncertainty: (i) the degree of basis-set convergence of the HF/aug-cc-pV5Z interaction energy, (ii) the accuracy of the aug-cc-pVTZ/aug-cc-pVQZ extrapolation of the MP2 cor-

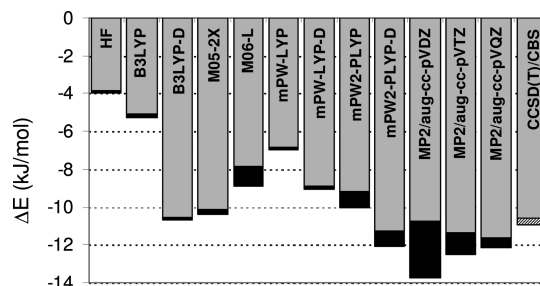


Figure 3. Ethyne methyl isocyanide interaction energies computed with different methods. Light gray: CP-corrected interaction energies. Black: BSSE. The sum of the black and gray fields equals the uncorrected interaction energy. The diagonally striped area shows the estimated uncertainty in the CCSD(T) interaction energy. All calculations were done using the uncorrected B3LYP/aug-cc-pVTZ optimized geometry. The aug-cc-pVTZ basis set was employed, unless stated otherwise.

relation energies, and (iii) the accuracy of the CCSD(T) correction term. Let us consider these three issues separately: (i) The HF interaction energies decrease by 0.73, 0.14, and 0.04 kJ/mol, from aug-cc-pVDZ to aug-cc-pVTZ, aug-cc-pVTZ to aug-cc-pVQZ, and aug-cc-pVQZ to aug-cc-pV5Z, respectively. Thus, the differences decrease by a factor of 5.6 and 3.0, respectively. Assuming this trend continues (i.e., the factor with which the differences decrease continues to halve), we expect the decrease from aug-cc-pV5Z to aug-cc-pV6Z to be $0.04/1.5 = 0.03 \text{ kJ/mol}$. We therefore estimate the HF/aug-cc-pV5Z interaction energy to be too large (i.e., too negative) by $0.03-0.05 \text{ kJ/mol}$. (ii) The accuracy of the MP2 extrapolation procedure can be deduced from the convergence of the MP2 correlation interaction energies. The correlation interaction energies decrease by 0.49 and 0.28 kJ/mol, from aug-cc-pVDZ to aug-cc-pVTZ and from aug-cc-pVTZ to aug-cc-pVQZ, respectively. Extrapolation to the CBS limit, using the aug-cc-pVTZ and aug-cc-pVQZ results, leads to a further decrease of 0.20 kJ/mol. The aug-cc-pVxZ ($x = D, T, Q$) MP2 interaction energies appear to converge smoothly toward the estimated CBS interaction energy, which therefore seems a realistic estimate of the true CBS limit. It seems reasonable to estimate the MP2/CBS limit to be accurate to $\sim 0.1 \text{ kJ/mol}$. It may appear unusual that the correlation interaction decreases when the basis set is enlarged. The reason for this is the use of uncorrected energies for the extrapolation, and the decrease in interaction reflects a reduced BSSE when more complete basis sets are employed. (iii) The CCSD(T) correction term differs by 0.15 kJ/mol for the aug-cc-pVDZ and aug-cc-pVTZ basis sets. This is in excellent agreement with Sinnokrot and Sherrill's results for different benzene dimer configurations, for which the CCSD(T) correction term also differed by less than 0.2 kJ/mol when evaluated with aug-cc-pVDZ or aug-cc-pVTZ.¹⁰⁰ As the aug-cc-pVTZ correction term should be more accurate than the aug-cc-pVDZ value, we estimate the aug-cc-pVTZ correction term to be accurate to $\sim 0.1 \text{ kJ/mol}$. In principle, we can avoid the correction term by using directly extrapolated CCSD(T) energies. Thus, we computed a second estimate of the CCSD(T)/CBS interaction energy, by extrapolation of the aug-cc-pVDZ and aug-cc-pVTZ

CCSD(T) correlation energies of the complex and the two fragments, using Halkier et al.'s extrapolation method⁹⁹ given in eq 4, and adding these to the HF/aug-cc-pV5Z total energies. The CCSD(T)/CBS interaction energy was then obtained by subtracting the thus obtained CCSD(T)/CBS monomer energies from the CCSD(T)/CBS dimer energy. This procedure yields a CCSD(T)/CBS value of -10.89 kJ/mol. Inspection of the aug-cc-pVDZ/aug-cc-pVTZ and aug-cc-pVTZ/aug-cc-pVQZ extrapolations of the MP2 results shows that the former yields an MP2/CBS interaction energy that is 0.28 kJ/mol larger than that obtained from the (more accurate) aug-cc-pVTZ/aug-cc-pVQZ extrapolation. Subtracting this extrapolation correction from the directly extrapolated CCSD(T)/CBS interaction energy yields a value of -10.62 kJ/mol, in excellent agreement with the CBS value obtained according to eq 3. Note that the CBS value obtained using directly extrapolated CCSD(T) energies differs from that obtained according to eq 3 only by the CCSD(T) correction term, taking the difference in the MP2 and CCSD(T) correlation energies extrapolated using the aug-cc-pVDZ and aug-cc-pVTZ basis sets instead of those computed with the aug-cc-pVTZ basis set. The good agreement between the two CBS values therefore confirms the small basis set dependence of the CCSD(T) correction term. Overall, we estimate the error in the estimated CCSD(T)/CBS interaction energy to be $+0.25/-0.15$ kJ/mol.

Figure 3 clearly demonstrates that the HF method and the B3LYP and mPW-LYP functionals severely underestimate the ethyne methyl isocyanide interaction energy. Also, the double hybrid mPW2-PLYP predicts too little stabilization. This is in agreement with previous work, where the mPW2-PLYP method was shown to underestimate the intramolecular dispersion in two conformers of the Tyr-Gly dipeptide.^{64,74} In addition, Benighaus et al. found that double hybrid functionals capture only about 50% of the interaction energy of dispersion-dominated complexes.⁸² The addition of an empirical dispersion term to the B3LYP, mPW-LYP, and mPW2-PLYP functionals brings the results in closer agreement with the CCSD(T) reference value, although mPW-LYP-D still underestimates, whereas mPW2-PLYP-D slightly overestimates the interaction energy. Note the significantly larger BSSE for the double hybrid functionals as compared to the standard DFT methods, which is caused by the added MP2 correlation term. The M05-2X functional gives a very good result, only slightly underestimating the CCSD(T) interaction energy. The CP-corrected MP2/aug-cc-pVDZ result is also very close to the CCSD(T) interaction energy. However, this is due to a cancellation of errors: basis set incompleteness vs MP2's tendency to overestimate dispersion interactions.^{100,101,110-112} When larger basis sets are used, the MP2 method overestimates the interaction energy, and the CCSD(T) correction term in eq 3 is therefore positive. Note also here the sizable BSSE, particularly when using the aug-cc-pVDZ basis set.

The CCSD(T)/CBS reference interaction energy was computed at the B3LYP/aug-cc-pVTZ optimized geometry and is therefore likely too small, as geometry relaxation at the CCSD(T) level is expected to increase the stability of the ethyne methyl isocyanide complex. Let us therefore return

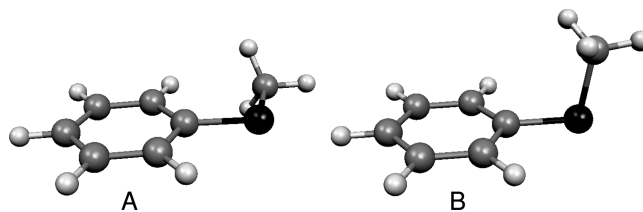


Figure 4. The planar (A) and perpendicular (B) conformations of thioanisole.

to the results displayed in Table 1, which were obtained using geometries optimized at the same level of theory as used for the energy calculation. The analysis above shows that MP2/aug-cc-pVTZ slightly overestimates, whereas M05-2X/aug-cc-pVTZ slightly underestimates. We therefore estimate the ethyne methyl isocyanide interaction energy to be 12 ± 1 kJ/mol.

3.2. Thioanisole. *3.2.1. B3LYP and MP2 Potential Energy Curves.* Previous computational and experimental work indicates that there are two possible configurations for the thioanisole molecule, with the thiomethyl group either planar or perpendicular with respect to the benzene ring (see Figure 4). The calculations done by Suzuki et al. showed that the global minimum geometry is drastically dependent on the level of theory used:²⁷ whereas B3LYP/cc-pVTZ yields a planar structure, MP2/6-311G(d,p) calculations result in the perpendicular conformer. MP2/cc-pVTZ finds both structures, with the planar one the most stable of these. The large basis set dependence of the MP2 results suggests that BSSE effects may be significant. We have therefore computed B3LYP and MP2 potential energy curves for rotation around the C(sp²)-S bond with and without CP corrections (Figure 5).

In agreement with Suzuki et al.,²⁷ the uncorrected MP2/6-311G(d,p) curve shows the perpendicular conformer at 90° as the only minimum. CP correction reveals a second, shallow minimum at 0° , though the perpendicular conformer remains the more stable one. All other curves show the planar conformer as the global minimum. The uncorrected MP2/cc-pVTZ curve shows a second, very shallow minimum at 90° (barely visible in Figure 5), which disappears after CP correction. The uncorrected and CP-corrected MP2/cc-pVTZ curves are very close to each other, even though the BSSE is fairly large in the MP2/cc-pVTZ calculations (~ 10 kJ/mol, see inset). However, the MP2/cc-pVTZ BSSE appears to be rather insensitive to the orientation of the thiomethyl group and therefore hardly affects the shape of the potential energy curve. The two B3LYP/6-311G(d,p) curves show a very shallow minimum around 80° . Such a secondary minimum was also found with some functionals in the study by Bossa et al.³⁴ With both basis sets, B3LYP predicts the relative stability of the perpendicular conformer to be larger as compared to MP2. As the potential energy curves appear to be rather sensitive to the basis set employed, we have investigated the basis set dependence in greater detail. To provide a reference to assess the convergence of the results obtained with different basis sets, we have constructed an estimated MP2/CBS and CCSD(T)/CBS profile. The CCSD(T)/CBS energies were obtained according to eq 5, whereas

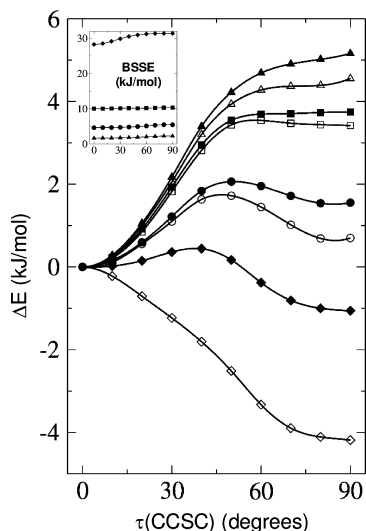


Figure 5. Potential energy curves for rotation around the $C(sp^2)$ -S bond computed with B3LYP/6-311G(d,p) (circles), B3LYP/cc-pVTZ (triangles), MP2/6-311G(d,p) (diamonds), and MP2/cc-pVTZ (squares). Open symbols denote uncorrected results; closed symbols denote CP-corrected results. The structures were optimized at fixed $\tau(\text{CCSC})$ torsion angles with M05-2X/6-31+G(d). The planar conformer ($\tau(\text{CCSC}) = 0^\circ$) was taken as the reference point for the relative energies. The inset shows the variation of the BSSE as a function of the $\tau(\text{CCSC})$ torsion angle.

the MP2/CBS energies were obtained in the same way, but without the CCSD(T) correction term.

Let us first discuss the accuracy of the CBS profiles. Here, we focus on the stability gap between the planar and perpendicular conformer (which for some, but not all, levels of theory corresponds to the rotational barrier). The stability gap is estimated to be 4.4 kJ/mol for MP2/CBS and 3.8 kJ/mol for CCSD(T)/CBS. As above, there are three main sources of uncertainty in this property: (i) the degree of basis-set convergence of the HF/cc-pV5Z energies, (ii) the accuracy of the aug-cc-pVTZ/aug-cc-pVQZ extrapolation of the MP2 correlation energies, and, for the CCSD(T)/CBS profile, (iii) the accuracy of the CCSD(T) correction term. (i) At the HF level, the stability gap decreases from 4.78 to 2.18, 1.71, and 1.53 kJ/mol for cc-pVDZ, cc-pVTZ, cc-pVQZ, and cc-pV5Z, and from 3.89 to 1.85 for aug-cc-pVDZ to aug-cc-pVTZ, respectively. On the basis of the convergence of these numbers, we estimate that the HF/cc-pV5Z stability gap would decrease by less than 0.2 kJ/mol upon a further increase of the basis set. (ii) The extrapolations of the MP2 correlation energy were performed both using the uncorrected and CP-corrected correlation energies. This should in principle lead to the same limits, as the BSSE vanishes at the CBS limit. Differences in the standard and CP-corrected CBS limits will therefore provide information on the accuracy of the extrapolation procedure. We focus again on the stability gap between the planar and perpendicular conformer. Using cc-pVTZ/cc-pVQZ extrapolation, this gap is 5.55 and 5.56 kJ/mol, for the standard and CP-corrected extrapolation, respectively, evidencing that CP correction does not have a noticeable effect on the cc-pVTZ/cc-pVQZ extrapolation. The uncorrected and CP-corrected

cc-pVTZ/cc-pVQZ extrapolated energies are almost identical over the whole $\tau(\text{CCSC})$ torsion angle range (Figure S1, Supporting Information). Using the less accurate cc-pVDZ/cc-pVTZ and aug-cc-pVDZ/aug-cc-pVTZ extrapolation, the differences between the standard and CP-corrected extrapolation are somewhat larger. Thus, CP-correction appears necessary when extrapolating using basis sets of DZ and TZ quality, and for such extrapolations we will only discuss the CBS limits obtained using the CP-corrected energies. The stability gap changes little from cc-pVTZ to cc-pVQZ (from 5.60 to 5.57 and from 5.48 to 5.53 kJ/mol, for the uncorrected and CP-corrected correlation energies), and the cc-pVTZ/cc-pVQZ-extrapolated CBS limit of 5.55/5.56 kJ/mol therefore seems entirely reasonable. The cc-pVDZ/cc-pVTZ CP-corrected extrapolated CBS limit is also in close agreement with the cc-pVTZ/cc-pVQZ-extrapolated limits. Aug-cc-pVDZ/aug-cc-pVTZ extrapolation using the CP-corrected correlation energies gives a somewhat lower CBS limit (5.25 kJ/mol) as compared to extrapolation using nonaugmented basis sets. However, the stability gap increases significantly (by 0.6 kJ/mol) from aug-cc-pVTZ to aug-cc-pVQZ, and the aug-cc-pVTZ/aug-cc-pVQZ extrapolation yields a CBS value of 5.9 kJ/mol, 0.3 kJ/mol larger than the CBS limits obtained using the unaugmented basis sets. There therefore is some uncertainty in the extrapolation of the MP2 correlation energies. The most accurate CBS limit is probably the one obtained with the largest basis sets employed (aug-cc-pVTZ/aug-cc-pVQZ), but the above analysis suggests that this CBS limit is more likely too large than too low. We therefore estimate the aug-cc-pVTZ/aug-cc-pVQZ extrapolation to be accurate to within $+0.2/-0.5$ kJ/mol. (iii) The last source of uncertainty in the CCSD(T)/CBS profile is related to the CCSD(T)/aug-cc-pVDZ correction term. On the basis of our analysis for the ethyne methyl isocyanide molecule above (section 3.1.3) and Sinnokrot and Sherrill's results for different benzene dimer configurations, we estimate that the CCSD(T)/aug-cc-pVDZ correction term is accurate to within 0.3 kJ/mol. Thus, the uncertainty in the MP2/CBS and CCSD(T)/CBS profiles is estimated to be $-0.7/+0.2$ kJ/mol and $-1.0/+0.5$ kJ/mol, respectively.

Let us now investigate the basis set dependence of the potential energy profiles. Figure 6 shows MP2 potential energy curves for rotation around the $C(sp^2)$ -S bond computed with different correlation consistent basis sets. The uncorrected cc-pVDZ results predict the perpendicular conformer to be more stable than the planar conformer. Counterpoise correction reverses this preference. All other basis sets predict the planar conformation as the global minimum. The results obtained with basis sets of triple- or quadruple- ζ quality, as well as the CCSD(T) reference profile, do not show a clear minimum or maximum at 90° ; instead the profile displays a plateau-like region around this torsion angle. Both CP correction and increasing the basis set size decreases the relative stability of the perpendicular conformer. Upon increasing the basis set quality, the uncorrected and CP-corrected curves converge toward each other. The CP-corrected cc-pVTZ and cc-pVQZ profiles and the uncorrected cc-pVQZ and aug-cc-pVQZ profiles are very close to each other. The aug-cc-pVTZ profiles, however,

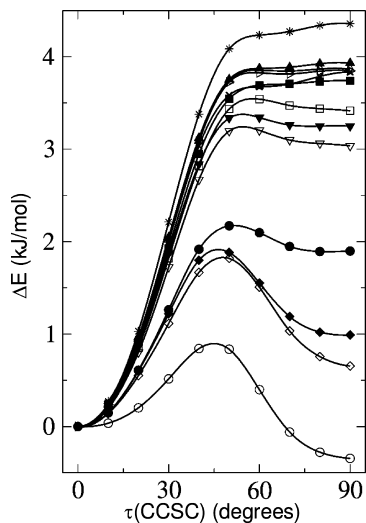


Figure 6. Potential energy curves for rotation around the $C(sp^2)$ –S bond computed with MP2 and various basis sets. Circles: cc-pVDZ. Squares: cc-pVTZ. Up triangles: cc-pVQZ. Diamonds: aug-cc-pVDZ. Down triangles: aug-cc-pVTZ. Right triangles: aug-cc-pVQZ. Open symbols denote uncorrected results; closed symbols denote CP-corrected results. The structures were optimized at fixed $\tau(\text{CCSC})$ torsion angles with M05-2X/6-31+G(d). The planar conformer ($\tau(\text{CCSC}) = 0^\circ$) was taken as the reference point for the relative energies. The estimated MP2/CBS potential energy curve is denoted by stars, whereas the estimated CCSD(T)/CBS potential energy curve is denoted by crosses.

show a smaller stability gap between the two thioanisole isomers. The estimated MP2/CBS profile, which is about 0.5 kJ/mol above the quadruple- ζ profiles at 90° , suggests that the calculations are not yet converged with respect to basis set quality, and basis sets beyond quadruple- ζ quality may be required to obtain converged results. Comparison with the CCSD(T)/CBS results suggests that, at the CBS limit, MP2 overestimates the relative stability of the planar conformer. However, the large uncertainty in the CBS profiles prevents solid conclusions on this.

Figure 7 shows the potential energy curves computed with B3LYP and the (aug)-cc-pVxZ basis sets of DZ and TZ quality. For the sake of clarity, the (aug)-cc-pVQZ results, which are close to the corresponding (aug)-cc-pVTZ values, are not shown. The QZ-quality results are included in the discussion on the convergence of the energy difference between the planar and perpendicular conformers (see section 3.2.2 below). All basis sets predict the planar conformer as the global minimum, whereas the perpendicular structure is not a minimum with any of the basis sets employed. The uncorrected B3LYP/cc-pVDZ shows a shallow minimum at 80° , which disappears after CP correction. All B3LYP curves (except the uncorrected cc-pVDZ curve) predict a larger stability gap between the planar and perpendicular configurations (barrier height 4–5 kJ/mol) as compared to CCSD(T)/CBS.

In summary, the thioanisole potential energy curves appear to be very basis set dependent, which also causes relatively large uncertainties in the estimated CBS profiles. When the MP2 method is used with small basis sets and no CP corrections, the perpendicular conformer is predicted to be

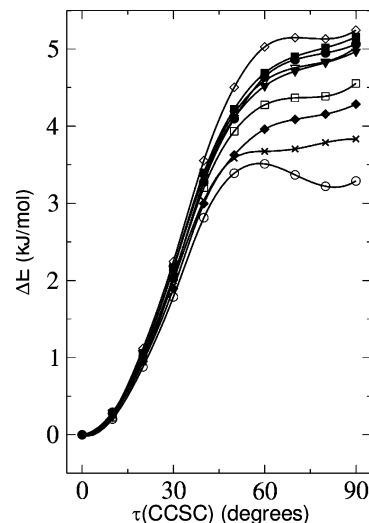


Figure 7. Potential energy curves for rotation around the $C(sp^2)$ –S bond computed with B3LYP and various basis sets. Circles: cc-pVDZ. Squares: cc-pVTZ. Diamonds: aug-cc-pVDZ. Down triangles: aug-cc-pVTZ. Open symbols denote uncorrected results; closed symbols denote CP-corrected results. The structures were optimized at fixed $\tau(\text{CCSC})$ torsion angles with M05-2X/6-31+G(d). The planar conformer ($\tau(\text{CCSC}) = 0^\circ$) was taken as the reference point for the relative energies. The estimated CCSD(T)/CBS potential energy curve is denoted by crosses.

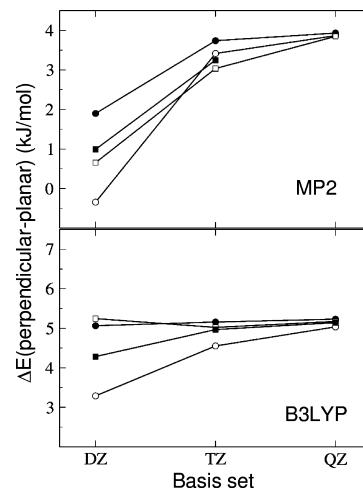


Figure 8. The energy gap between the planar and perpendicular conformers as a function of basis set size calculated with MP2 (upper plot) and B3LYP (lower plot). Circles: cc-pVxZ. Squares: aug-cc-pVxZ. Open symbols denote uncorrected results; closed symbols denote CP-corrected results.

the global minimum. Larger basis sets, however, show the planar conformer as the global minimum. MP2 appears to overestimate the energy gap between the planar and perpendicular configurations. B3LYP predicts the global minimum to be planar, irrespective of the basis set employed. However, like MP2, B3LYP overestimates the relative stability of this conformer.

3.2.2. Convergence of the Energy Difference between the Planar and Perpendicular Configurations. Figure 8 shows the B3LYP and MP2 energy gap between the planar and perpendicular conformers as a function of basis set

quality. Both uncorrected as well as CP-corrected results are shown. All MP2 curves and all B3LYP curves clearly converge toward a limiting value. The limiting value is 4.4 kJ/mol for MP2 (see above) and is estimated from Figure 8 to be about 5.2 kJ/mol for B3LYP. Both limits are higher than the CCSD(T) estimate of 3.8 kJ/mol.

The B3LYP curves vary less with basis set improvement than the MP2 curves (note that the energy range on the y axis is identical for the two plots). This is consistent with the smaller basis set dependence of DFT as compared to MP2. CP correction brings the results closer to the limiting value, except for the B3LYP/aug-cc-pVDZ calculation. However, CP correction does restore the regular convergence of the B3LYP/aug-cc-pVxZ curve. This is consistent with earlier work that showed that, in many cases, the convergence behavior of molecular properties calculated with the correlation consistent basis sets is significantly improved if BSSE is taken into account.¹¹³ This agreement indicates that the simple CP scheme to correct for the intramolecular BSSE used in the current study works well.

Even though the BSSE is significantly smaller in the B3LYP calculations (ranging from 6.8–8.6 kJ/mol for B3LYP/cc-pVDZ compared to 22.4–24.6 kJ/mol for MP2/cc-pVDZ), the effect on the B3LYP/cc-pVDZ and MP2/cc-pVDZ results is similar (B3LYP, 1.8 kJ/mol; MP2, 2.2 kJ/mol). CP correction of the B3LYP/cc-pVDZ calculation brings the energy gap in close proximity to the limiting value.

The corresponding uncorrected and CP-corrected curves using the aug-cc-pVxZ basis sets are closer to each other than those computed with the unaugmented basis sets. This is particularly evident for the MP2 method. This does not mean that the BSSE is smaller for the augmented basis sets. In fact, the BSSE has a similar magnitude for the corresponding standard and augmented basis sets. The close agreement between the uncorrected and CP-corrected aug-cc-pVxZ curves results from the fact that the magnitude of the BSSE is very similar for the planar and perpendicular conformers in the aug-cc-pVxZ calculations.

3.2.3. Comparison of Different Electronic Structure Methods. Figure 9 shows the potential energy curves for rotation around the C(sp²)–S bond computed with different electronic structure methods, the aug-cc-pVDZ basis set, and no CP corrections.

The HF curve shows the perpendicular structure as the global minimum, in agreement with previous studies,^{28–30} with only a very shallow minimum at 0°. The curve is similar to the uncorrected MP2/6-31G(d,p) profile (Figure 5). However, unlike for MP2, increasing the basis set does not result in the planar structure becoming the global minimum, even though the energy gap between the two conformations reduces somewhat (Figure S2, Supporting Information). CCSD(T)/aug-cc-pVDZ predicts only a very small stability difference (0.13 kJ/mol) between the planar and perpendicular conformer. At the CBS limit, however, the planar conformer is 3.83 kJ/mol more stable than the perpendicular structure. The MP2/aug-cc-pVDZ stability difference between the two conformers is also small (0.65 kJ/mol), but we have seen above that, at the CBS limit, MP2 overestimates the stability gap. As we have seen above, B3LYP

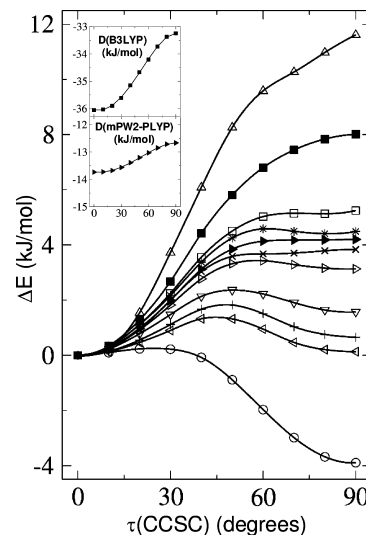


Figure 9. Potential energy curves for rotation around the C(sp²)–S bond computed at different levels of theory. All profiles (except the CCSD(T)/CBS reference curve) were computed with the aug-cc-pVDZ basis set. Open circles: HF. Left triangles: CCSD(T). Pluses: MP2. Down triangles: mPW-LYP. Open right triangles: mPW2-PLYP. Closed right triangles: mPW2-PLYP-D. Open squares: B3LYP. Closed squares: B3LYP-D. Up triangles: M06-L. Stars: M05-2X. The inset shows the van der Waals/dispersion corrections (D) computed by the B3LYP-D and mPW2-PLYP-D methods. The structures were optimized at fixed $\tau(\text{CCSC})$ torsion angles with M05-2X/6-31+G(d). The planar conformer ($\tau(\text{CCSC}) = 0^\circ$) was taken as the reference point for the relative energies. The estimated CCSD(T)/CBS potential energy curve is denoted by crosses.

predicts a too large stability gap between the two conformers. The addition of a dispersion term exacerbates this (B3LYP-D profile). The inset clearly shows the dispersion contribution's preference for the planar conformation. Because the dispersion term, as calculated in the Orca program, is by default independent of the basis set employed, B3LYP-D exhibits the same basis set dependence as B3LYP. It can thus be inferred from Figure 7 that the use of larger basis sets in the B3LYP-D calculations (or CP-correction) would only slightly decrease the discrepancy with the CCSD(T) curve, but the B3LYP-D stability gap would remain significantly too large. Interestingly, other systems for which DFT-D was found to give results in disagreement with CCSD(T) also include an anisole unit: anisole–water and anisole–ammonium.⁷⁹ For these systems, the overestimated interaction energies obtained by DFT-D were attributed to an overestimated dispersion correction or to double counting of electron correlation effects by the DFT and van der Waals parts of the method. The mPW2-PLYP/aug-cc-pVDZ and mPW2-PLYP-D/aug-cc-pVDZ profiles are very close to the reference CCSD(T)/CBS curve and also show the plateau region around 90°. Increasing the basis set quality in the mPW2-PLYP-D calculations deteriorates the results somewhat (Figure S3, Supporting Information). Note that the dispersion contribution is smaller for mPW2-PLYP than for B3LYP (inset), as the nonlocal perturbation term in the double hybrid functional already accounts for part of the dispersion energy. The M05-2X/aug-cc-pVDZ profile is also close to the reference profile.

Like the mPW2-PLYP results, the M05-2X profile is very basis set dependent (Figure S4, Supporting Information), and basis sets of at least aug-cc-pVDZ or cc-pVTZ quality are required to obtain good agreement with the CCSD(T)/CBS curve. Here, we also investigated the performance of the aug-cc-pV($n+d$)Z basis sets,¹⁰⁷ which contain additional tight d functions for second-row elements. Figure S4 shows that these basis sets give only slightly revised profiles compared to those obtained with basis sets without the tight d-functions. M06-L significantly overestimates the stability gap between the planar and perpendicular conformation. Improving the basis set and correcting for BSSE does not change the M06-L profile significantly (Figure S5, Supporting Information).

4. Summary and Conclusions

We have investigated two molecular systems, the ethyne methyl isocyanide complex and thioanisole; for both molecules, previous studies had shown significant discrepancies between the results obtained with the MP2 and B3LYP methods. The ethyne methyl isocyanide complex exhibits two π bonds. Earlier work by Cao et al.²⁶ showed that B3LYP/cc-pVTZ predicts significantly longer H-bond distances than MP2/cc-pVTZ. In the current work, we show that this is likely due to missing dispersion in the B3LYP calculations. Comparison with estimated CCSD(T)/CBS results shows that B3LYP significantly underestimates the stability of the complex, whereas MP2 slightly overestimates (when basis sets larger than aug-cc-pVDZ are employed). Other methods that give results in close proximity to the CCSD(T)/CBS reference value include M05-2X, B3LYP-D, and (CP-corrected) mPW2-PLYP-D.

The thioanisole molecule can adopt two different structures, with the thiomethyl group planar or perpendicular with respect to the benzene ring. An earlier study by Suzuki et al.²⁷ showed that B3LYP/cc-pVTZ geometry optimizations yield the planar conformer, whereas MP2/6-311G(d,p) geometry optimizations lead to the perpendicular conformer. MP2/cc-pVTZ shows both structures as a minimum, with the planar conformer as the global minimum. By comparison with estimated CCSD(T)/CBS results, we show in this work that the preferred thioanisole structure is planar. The energy landscape is very shallow around the perpendicular configuration, and the CCSD(T)/CBS energy profile for rotation around the C(sp²)-S bond shows a plateau region, rather than a true minimum or maximum, near the perpendicular configuration. The energy barrier for rotation around the C(sp²)-S bond is estimated to be 3.8 -1.0/+0.5 kJ/mol. This small energy barrier indicates nearly free rotation around this bond, which may explain the varied conformations found experimentally for thioanisole. The stability gap between the planar and perpendicular conformations is found to be very method and basis set dependent. Thus, the 6-311G(d,p) basis set is too small to give the correct conformer in MP2 calculations. Both B3LYP (with basis sets of at least triple- ζ quality) and MP2 (at the CBS limit) appear to slightly overestimate the stability gap between the two thioanisole conformations. B3LYP predicts a maximum instead of a plateau region at the perpendicular structure. This is presumably not due to missing dispersion in the B3LYP calculations,

as the addition of an empirical dispersion term further disfavors the perpendicular conformer. Other density functionals, like M06-L and M05-2X, also show a maximum in the profile at the perpendicular conformer. Note that the uncertainty in the estimated CCSD(T)/CBS limit is relatively large, due to the large basis set dependence of the results. Results that are within the estimated error bars of the CCSD(T)/CBS reference value of the stability gap between the planar and perpendicular conformer include MP2 with basis sets of at least triple- ζ quality, B3LYP with cc-pVDZ or CP-corrected aug-cc-pVDZ, M05-2X/cc-pVTZ, mPW2-PLYP/aug-cc-pVDZ, and mPW2-PLYP-D/(aug-)cc-pVDZ. B3LYP calculations with basis sets larger than aug-cc-pVDZ and mPW2-PLYP-D calculations with basis sets of at least triple- ζ quality yield overestimated values of the stability gap. M05-2X calculations with basis sets smaller than cc-pVTZ underestimate, whereas larger basis sets overestimate the stability gap.

The ethyne methyl isocyanide complex shows the typical characteristics of a dispersion-dominated system: B3LYP underestimates the interaction; MP2 slightly overestimates; and methods like M05-2X, B3LYP-D, and mPW2-PLYP-D give good results. The situation is different however for the thioanisole molecule. The energy profile for rotation around the C(sp²)-S bond is very method and basis set dependent, and methods that could be expected to yield good results, like B3LYP-D, do not agree with the reference profile. Surprising is also the vastly different results obtained with M05-2X and M06-L, two basis sets taken from the related M05 and M06 density functional families of the Minnesota group. One possible explanation for the difficulty in describing this system may be the presence of the sulfur atom. Indeed, previous studies indicate that the rotational profile is less method-dependent for anisole as compared to thioanisole.^{30,34} We have not systematically explored the addition of tight d functions, which were found to improve the convergence of the results for molecules with second-row atoms.¹⁰⁷ However, exploratory M05-2X and M06-L calculations using the (aug-)cc-pV($n+d$)Z ($n = 2, 3$) basis sets indicate that the tight d functions only have a moderate influence upon the results.

For both systems, the effects of BSSE were investigated. The standard CP method was used to correct the *intermolecular* BSSE in the ethyne methyl isocyanide complex as well as the *intramolecular* BSSE in thioanisole. For ethyne methyl isocyanide, it is found that CP-corrected geometry optimization is not required when basis sets of at least triple- ζ quality are used. However, as BSSE increases the MP2/aug-cc-pVTZ interaction energy by 10%, for accurate results the interaction energies should be corrected for BSSE. The intramolecular BSSE in thioanisole affects the relative stability of the planar and perpendicular conformers at both the MP2 and B3LYP levels of theory, particularly when basis sets of less than triple- ζ quality are employed. In general, CP correction brings the result closer to the limiting value for the method employed. Though this is not the case for the B3LYP/aug-cc-pVDZ calculation, CP correction does improve the convergence behavior of the energy gap computed with B3LYP/aug-cc-pV x Z. Thus, the simple CP

scheme used here to correct the intramolecular BSSE in thioanisole appears to work well.

Acknowledgment. We gratefully acknowledge EaSt-CHEM for computational support via the EaStCHEM Research Computing Facility.

Supporting Information Available: Plots containing potential energy curves for rotation around the C(sp²)-S bond computed with various methods and basis sets. Cartesian coordinates of the ethyne methyl isocyanide complex and the ethyne and methyl isocyanide monomers optimized at different levels of theory. Cartesian coordinates of thioanisole optimized with M05-2X/6-31+G(d) at fixed τ (CCSC) torsion angles. This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- Burley, S. K.; Petsko, G. A. *Science* **1985**, *229*, 23–28.
- Saenger, W. *Principles of Nucleic Acid Structure*; Springer-Verlag: New York, 1984.
- Meyer, E. A.; Castellano, R. K.; Diederich, F. *Angew. Chem., Int. Ed.* **2003**, *42*, 1210–1250.
- Lin, W.; Evans, O. R.; Xiong, R.-G.; Wang, Z. *J. Am. Chem. Soc.* **1998**.
- Zhao, R.; Matsumoto, S.; Akazome, M.; Ogura, K. *Tetrahedron* **2002**, *58*, 10233–10241.
- Møller, C.; Plesset, M. S. *Phys. Rev.* **1934**, *46*, 618–622.
- Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648–5652.
- Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785–789.
- Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098–3100.
- Kristyán, S.; Pulay, P. *Chem. Phys. Lett.* **1994**, *229*, 175–180.
- Hobza, P.; Šponer, J.; Reschel, T. *J. Comput. Chem.* **1995**, *16*, 1315–1325.
- Pérez-Jordá, J. M.; Becke, A. D. *Chem. Phys. Lett.* **1995**, *233*, 134–137.
- Tsuzuki, S.; Uchimaru, T.; Tanabe, K. *Chem. Phys. Lett.* **1998**, *287*, 202–208.
- Kohn, W.; Meir, Y.; Makarov, D. E. *Phys. Rev. Lett.* **1998**, *80*, 4153–4156.
- Millet, A.; Korona, T.; Moszynski, R.; Kochanski, E. *J. Chem. Phys.* **1999**, *111*, 7727–7735.
- Rappé, A. K.; Bernstein, E. R. *J. Phys. Chem. A* **2000**, *104*, 6117–6128.
- Kurita, N.; Sekino, H. *Chem. Phys. Lett.* **2001**, *348*, 139–146.
- Tsuzuki, S.; Lüthi, H. P. *J. Chem. Phys.* **2001**, *114*, 3949–3957.
- van Mourik, T.; Gdanitz, R. J. *J. Chem. Phys.* **2002**, *116*, 9620–9623.
- Johnson, E. R.; Wolkow, R. A.; DiLabio, G. A. *Chem. Phys. Lett.* **2004**, *394*, 334–338.
- Cybulski, S. M.; Seversen, C. E. *J. Chem. Phys.* **2005**, *122*, 014117.
- Dąbkowska, I.; Jurecka, P.; Hobza, P. *J. Chem. Phys.* **2005**, *122*, 204322.
- van Mourik, T. *Chem. Phys.* **2004**, *304*, 317–319.
- Holroyd, L. F.; van Mourik, T. *Chem. Phys. Lett.* **2007**, *442*, 42–46.
- Shields, A. E.; van Mourik, T. *J. Phys. Chem. A* **2007**, *111*, 13272–13277.
- Cao, D.; Ren, F.; Feng, X.; Wang, J.; Li, Y.; Hu, Z.; Chen, S. *THEOCHEM* **2008**, *849*, 76–83.
- Nagasaka-Hoshino, M.; Isozaki, T.; Suzuki, T.; Ichimura, T.; Kawauchi, S. *Chem. Phys. Lett.* **2008**, *457*, 58–61.
- Vondrák, T.; Sato, S.; Špirko, V.; Kimura, K. *J. Phys. Chem. A* **1997**, *101*, 8631–8638.
- Dal Colle, M.; Giuseppe Distefano, G.; Jones, D.; Modelli, A. *J. Phys. Chem. A* **2000**, *104*, 8227–8235.
- Dolgounitcheva, O.; Zakrzewski, V. G.; Ortiz, J. V. *Int. J. Quantum Chem.* **1998**, *70*, 1037–1043.
- Gellini, V.; Moroni, L.; Muniz-Miranda, M. *J. Phys. Chem. A* **2002**, *106*, 10999–11007.
- Bzhezovskii, V. M.; Kapustin, E. G. *Russ. J. Org. Chem.* **2002**, *38*, 564–572.
- Bzhezovskii, V. M.; Kapustin, E. G. *Russ. J. Gen. Chem.* **2004**, 74–82.
- Bossa, M.; Morpurgo, S.; Stranges, S. *THEOCHEM* **2002**, *618*, 155–164.
- Shishkov, I. F.; Khristenko, L. V.; Karasev, N. M.; Vilkov, L. V.; Oberhammer, H. *J. Mol. Struct.* **2008**, *873*, 137–141.
- Dewar, P. S.; Ernstbrunner, E.; Gilmore, J. R.; Godfrey, M.; Mellor, J. M. *Tetrahedron* **1974**, *30*, 2455–2459.
- Zaripov, N. M. *J. Struct. Chem.* **1976**, *17*, 640–642.
- Schweig, A.; Thon, N. *Chem. Phys. Lett.* **1976**, *38*, 482–485.
- Jones, I. W.; Tebby, J. C. *J. Chem. Soc., Perkin Trans.* **1979**, *2*, 217–218.
- Honegger, E.; Heilbronner, E. *Chem. Phys. Lett.* **1981**, *81*, 615–619.
- Mohraz, M.; Jian-qi, J.; Heilbronner, E.; Solladiéa-Cavallo, A.; Matloubi-Moghadam, F. *Helv. Chim. Acta* **1981**, *64*, 97–112.
- Emsley, J. W.; Longeri, M.; Veracini, C. A.; Catalano, D.; Pedulli, G. F. *J. Chem. Soc., Perkin Trans.* **1982**, *2*, 1289–1296.
- Lumbroso, H.; Liégeois, C.; Testaferri, L.; Tiecco, M. *J. Mol. Struct.* **1986**, *144*, 121–133.
- Schaefer, T.; Baleja, J. D. *Can. J. Chem.* **1986**, *64*, 1326–1331.
- Schaefer, T.; Penner, G. H. *Can. J. Chem.* **1988**, *66*, 1229–1238.
- Celebre, G.; Longeri, M.; Emsley, J. W. *Appl. J. Magn. Res.* **1991**, *2*, 611–625.
- Schaefer, T.; Sebastian, R.; Salman, S. R.; Baleja, J. D.; Penner, G. H.; McKinnon, D. M. *Can. J. Chem.* **1991**, *69*, 620–624.
- Chmielewski, D.; Werstiuk, N. H.; Wildman, T. A. *Can. J. Chem.* **1993**, *71*, 1741–1750.

- (49) Bzhezovsky, V. M.; Penkovsky, V. V.; Rozhenko, A. B.; Iksanova, S. V.; Kondratenko, N. V.; Yagupolsky, L. M. *J. Fluorine Chem.* **1994**, *69*, 41–48.
- (50) Yamakita, Y.; Isogai, Y.; Ohno, K. *J. Chem. Phys.* **2006**, *124*, 104301.
- (51) Bakó, I.; Megyes, T.; Pálinkás, G. *Chem. Phys.* **2005**, *316*, 235–244.
- (52) Hopkins, B. W.; ElSohly, A. M.; Tschumper, G. S. *Phys. Chem. Chem. Phys.* **2007**, *9*, 1550–1558.
- (53) Hopkins, B. W.; Tschumper, G. S. *J. Phys. Chem. A* **2004**, *108*, 2941–2948.
- (54) Hopkins, B. W.; Tschumper, G. S. *Chem. Phys. Lett.* **2005**, *407*, 362–367.
- (55) Zhao, Y.; Truhlar, D. G. *J. Chem. Phys.* **2006**, *125*, 194101.
- (56) Zhao, Y.; Schultz, N. E.; Truhlar, D. G. *J. Chem. Theory Comput.* **2006**, *2*, 364–382.
- (57) Grimme, S. *J. Chem. Phys.* **2006**, *124*, 034108.
- (58) Grimme, S.; Schwabe, T. *Phys. Chem. Chem. Phys.* **2006**, *8*, 4398–4401.
- (59) Grimme, S. *J. Comput. Chem.* **2004**, *25*, 1463–1473.
- (60) Grimme, S. *J. Comput. Chem.* **2006**, *27*, 1787–1799.
- (61) Zhao, Y.; Truhlar, D. G. *J. Chem. Theory Comput.* **2007**, *3*, 289–300.
- (62) Sousa, S. F.; Fernandes, P. A.; Ramos, M. J. *J. Phys. Chem. A* **2007**, *111*, 10439–10452.
- (63) Zhao, Y.; Truhlar, D. G. *J. Am. Chem. Soc.* **2007**, *129*, 8440–8442.
- (64) van Mourik, T. *J. Chem. Theory Comput.* **2008**, *4*, 1610–1619.
- (65) Leverentz, H. R.; Truhlar, D. G. *J. Phys. Chem. A* **2008**, *112*, 6009–6016.
- (66) Dahlke, E. E.; Olson, R. M.; Leverentz, H. R.; Truhlar, D. G. *J. Phys. Chem. A* **2008**, *112*, 3976–3984.
- (67) Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. C* **2008**, *112*, 6860–6868.
- (68) Zhao, Y.; Truhlar, D. G. *Acc. Chem. Res.* **2008**, *41*, 157–167.
- (69) Benitez, D.; Tkatchouk, E.; Yoon, I. I.; Fraser Stoddart, J.; Goddard, W. A., III. *J. Am. Chem. Soc.* **2008**, *130*, 14928–14929.
- (70) Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2008**, *112*, 6794–6799.
- (71) Zhao, Y.; Truhlar, D. G. *Theor. Chem. Acc.* **2008**, *120*, 215–241.
- (72) Valero, R.; Costa, R.; Moreira, I. D. P. R.; Truhlar, D. G.; Illas, F. *J. Chem. Phys.* **2008**, *128*, 114103.
- (73) Bryantsev, V. S.; Diallo, M. S.; van Duin, A. C. T.; Goddard, W. A., III. *J. Chem. Theory Comput.* **2009**, *5*, 1016–1026.
- (74) Cao, J.; van Mourik, T. *Chem. Phys. Lett.* **2010**, *485*, 40–44.
- (75) Antony, J.; Grimme, S. *Phys. Chem. Chem. Phys.* **2006**, *8*, 5287–5293.
- (76) Piacenza, M.; Grimme, S. *J. Am. Chem. Soc.* **2005**, *127*, 14841–14848.
- (77) Piacenza, M.; Grimme, S. *ChemPhysChem* **2005**, *6*, 1554–1558.
- (78) Morgado, V.; M.A., V.; Hillier, I. H.; Shan, X. *Phys. Chem. Chem. Phys.* **2007**, *9*, 448–451.
- (79) Barone, V.; Biczyński, M.; Pavone, M. *Chem. Phys.* **2008**, *346*, 247–256.
- (80) Tarnopolsky, A.; Karton, A.; Sertchook, R.; Vuzman, D.; Martin, J. M. L. *J. Phys. Chem. A* **2008**, *112*, 3–8.
- (81) Schwabe, T.; Grimme, S. *Phys. Chem. Chem. Phys.* **2007**, *9*, 3397–3406.
- (82) Benighaus, T.; DiStasio, R. A., Jr.; Lochan, R. C.; Chai, J.-D.; Head-Gordon, M. *J. Phys. Chem. A* **2008**, *112*, 2702–2712.
- (83) Seal, P.; Chakrabarti, S. *J. Phys. Chem. A* **2009**, *113*, 1377–1383.
- (84) Boys, S. F.; Bernardi, F. *Mol. Phys.* **1970**, *19*, 553–566.
- (85) van Mourik, T.; Karamertzanis, P. G.; Price, S. L. *J. Phys. Chem. A* **2006**, *110*, 8–12.
- (86) Valdés, H.; Klusák, V.; Pitoňák, M.; Exner, O.; Starý, I.; Hobza, P.; Rulíšek, L. *J. Comput. Chem.* **2008**, *29*, 861–870.
- (87) Wilson, A. K.; van Mourik, T.; Dunning, T. H., Jr. *THEOCHEM* **1996**, *388*, 339–349.
- (88) van Mourik, T.; Dunning, T. H., Jr. *Int. J. Quantum Chem.* **2000**, *76*, 205–221.
- (89) Palermo, N. Y.; Csontos, J.; Owen, M. C.; Murphy, R. F.; Lovas, S. *J. Comput. Chem.* **2007**, *28*, 1208–1214.
- (90) van Mourik, T. *J. Comput. Chem.* **2008**, *29*, 103.
- (91) Jensen, F. *Chem. Phys. Lett.* **1996**, *261*, 633–636.
- (92) Jensen, F. *J. Chem. Theory Comput.* **2010**, *6*, 100–106.
- (93) Balabin, R. M. *J. Chem. Phys.* **2010**, 211103.
- (94) Asturiol, D.; Duran, M.; Salvador, P. *J. Chem. Phys.* **2008**, *128*, 144108.
- (95) Asturiol, D.; Duran, M.; Salvador, P. *J. Chem. Theory Comput.* **2009**, *5*, 2574–2581.
- (96) Balabin, R. M. *J. Chem. Phys.* **2008**, *129*, 164202.
- (97) Klopper, W.; Lüthi, H. P. *Mol. Phys.* **1999**, *96*, 559–570.
- (98) Jurečka, P.; Šponer, J.; Černý, J.; Hobza, P. *Phys. Chem. Chem. Phys.* **2006**, *8*, 1985–1993.
- (99) Halkier, A.; Helgaker, T.; Jørgensen, P.; Klopper, W.; Koch, H.; Olsen, J.; Wilson, A. K. *Chem. Phys. Lett.* **1998**, *286*, 243–252.
- (100) Sinnokrot, M. O.; Sherrill, C. D. *J. Phys. Chem. A* **2004**, *108*, 10200–10207.
- (101) Tsuzuki, S.; Honda, K.; Uchimaru, T.; Mikami, M.; Tanabe, K. *J. Am. Chem. Soc.* **2002**, *124*, 104–112.
- (102) Jurečka, P.; Hobza, P. *Chem. Phys. Lett.* **2002**, *365*, 89–94.
- (103) 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.; 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.
- (104) Neese, F. *ORCA - an ab initio, density functional and semiempirical program package*, 2.6, revision 35; University of Bonn: Bonn, Germany, 2007.
- (105) MOLPRO (2002.10) is a package of ab initio programs written by Werner, H.-J. and Knowles, P. J. The authors are Werner, H.-J.; Knowles, P. J.; Schütz, M.; Lindh, R.; 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.; Manby, F. R.; Meyer, W.; Mura, M. E.; Nicklaß, A.; Palmieri, P.; Pitzer, R.; Schumann, U.; Stoll, H.; Stone, A. J.; Tarroni, R.; Thorsteinsson, T.
- (106) Vosko, S. H.; Wilk, L.; Nusair, M. *Can. J. Phys.* **1980**, *58*, 1200–1211.
- (107) Dunning, T. H., Jr; Peterson, K. A.; Wilson, A. K. *J. Chem. Phys.* **2001**, *114*, 9244–9253.
- (108) Feller, D. *J. Comput. Chem.* **1996**, *17*, 1571–1586.
- (109) Schuchardt, K. L.; Didier, B. T.; Elsethagen, T.; Sun, L.; Gurumoorthi, V.; Chase, J.; Li, J.; Windus, T. L. *J. Chem. Inf. Model.* **2007**, *47*, 1045–1052.
- (110) Sinnokrot, M. O.; Sherrill, C. D. *J. Am. Chem. Soc.* **2004**, *126*, 7690–7697.
- (111) Kolář, M.; Hobza, P. *J. Phys. Chem. A* **2007**, *111*, 5851–5854.
- (112) Pitoňák, M.; Riley, K. E.; Neogrády, P.; Hobza, P. *ChemPhysChem* **2008**, *9*, 1636–1644.
- (113) van Mourik, T.; Wilson, A. K.; Peterson, K. A.; Woon, D. E.; Dunning, T. H., Jr. *Adv. Quantum Chem.* **1999**, *31*, 105–135.

CT100295F

JCTC

Journal of Chemical Theory and Computation

Scalar and Spin–Orbit Relativistic Corrections to the NICS and the Induced Magnetic Field: The case of the E_{12}^{2-} Spherenes (E = Ge, Sn, Pb)

Abril Carolina Castro,[†] Edison Osorio,[‡] J. Oscar C. Jiménez-Halla,[†] Eduard Matito,[§]
William Tiznado,^{*,‡} and Gabriel Merino^{*,†}

Departamento de Química, División de Ciencias Naturales y Exactas, Universidad de Guanajuato, Col. Noria Alta s/n C.P. 36050, Guanajuato, Gto, México, Departamento de Ciencias Químicas, Facultad de Ecología y Recursos Naturales, Universidad Andres Bello, Av. República 275, Santiago-Chile, and Institute of Physics, University of Szczecin, 70-451 Szczecin, Poland

Received June 5, 2010

Abstract: Can relativistic effects modify the NICS and the \mathbf{B}^{ind} values? In this manuscript we evaluate the relativistic corrections incorporated via the zeroth-order regular approximation to the calculations of nucleus-independent chemical shifts and the induced magnetic field (\mathbf{B}^{ind}) in the E_{12}^{2-} spherenes (E = Ge, Sn, Pb). We found that both electron delocalization descriptors are strongly affected by the relativistic corrections. For instance, for plumbaspherene, the difference in values from the nonrelativistic to the relativity-included calculation is almost 40 ppm! Our results show that the changes observed in the NICS and \mathbf{B}^{ind} values in the title cages are a consequence of the treatment of the relativistic effects. If these effects are included as scalar or spin–orbit calculations, then we can establish the next trend: Ge_{12}^{2-} is a nonaromatic species, Sn_{12}^{2-} is a low aromatic species, and Pb_{12}^{2-} is strongly aromatic, according to calculated NICS and \mathbf{B}^{ind} values. Thus, any prediction of electron delocalization in molecules containing heavy elements without considering an adequate treatment for relativistic effects may lead to an erroneous chemical interpretation.

The discovery of fullerene has motivated the quest for other stable spherical clusters. A spherical molecule, what could be more perfect? No other analogous gas-phase clusters were found or yielded to bulk syntheses until 2007.¹ In that year, Cui et al. found that both stannaspherene and plumbaspherene are stable clusters with a delocalized spherical π -bonding, similar to buckminsterfullerene C_{60} .^{2,3} Stannaspherene and plumbaspherene are clusters formed by 12 Sn and Pb atoms, respectively, distributed in a perfect icosahedron. Both cages have diameters comparable to that of C_{60} and can be considered as inorganic analogues of the buckyball. The large internal space in such beautiful systems

can trap some transition-metal atom to form new endohedral cage clusters, $\text{M}@E_{12}$, analogous to endohedral fullerenes.^{4–12} Recently, Sun et al. reported that during the attempt to synthesize endohedral stannaspherenes, they crystallized a new $\text{Pd}_2@\text{Sn}_{18}^{4-}$ cluster,¹⁰ which can be viewed as the fusion of two $\text{Pd}@\text{Sn}_{12}^{2-}$ clusters, suggesting that a large number of endohedrally doped species can be synthesized in the bulk using the title systems.

Particularly for Pb_{12}^{2-} , it is considered that the large stability of the endohedral $\text{M}@\text{Pb}_{12}$ clusters is a delicate balance between the cavity size and aromaticity.¹³ In this context, Chen et al. calculated the nucleus-independent chemical shift (NICS) value at the Sn_{12}^{2-} cage center (2.5 ppm at the GIAO-B3LYP/LanL2DZdp//B3LYP/LanL2DZdp level), indicating a nonaromatic character of stannospherene.⁵ According to this result, it is not possible to associate large stability to aromaticity in this species. Recently, Tian and

* Corresponding authors. E-mail: gmerino@quijote.ugto.mx; wtiznado@unab.cl.

[†] Universidad de Guanajuato.

[‡] Universidad Andres Bello.

[§] University of Szczecin.

co-workers evaluated the NICS at the same position and obtained a value of -5.0 ppm (calculated at the B3LYP/aug-cc-pVDZ-PP level).¹⁴ They suggested that a large basis set, including d-orbitals, is needed to analyze aromaticity in the aforementioned compound, thus the presence of a d-orbital is essential to calculate correctly the NICS in Sn_{12}^{2-} . However, they did not discuss the role of the relativistic effects, which can also drastically affect the shielding tensor,^{15–22} and the descriptors derived from it, such as the NICS.^{23–26}

In this manuscript we evaluate the relativistic corrections incorporated via the zeroth-order regular approximation (ZORA) to the calculations of NICS^{27,28} and the induced magnetic field (\mathbf{B}^{ind}).^{29,30} We found that in the case of Sn_{12}^{2-} both electron delocalization descriptors are strongly affected by the relativistic corrections. We also include the Ge_{12}^{2-} and Pb_{12}^{2-} clusters in this study in order to compare the magnitude of these changes.

Computational Details

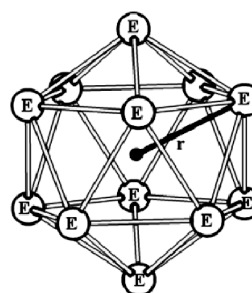
The geometries of these molecules have been optimized using the exchange functional of Becke with the correlation functional of Perdew (this combination is the so-called BP86 density functional).^{31,32} Uncontracted Slater-type orbitals (STOs) were employed as the basis functions for the self-consistent field (SCF) calculations. The basis sets have triple- ζ quality augmented by two sets of polarization functions (TZ2P), that is, d and f functions for the all atoms. In addition, we performed harmonic frequency calculations in order to check for the correct assessment of the minima points studied. The shielding tensors were calculated at the BP86, PW91,³³ and SAOP^{34,35} levels in conjunction with a TZ2P basis set. Both the scalar and the spin-orbit (SO) relativistic effects on the geometry and the shielding tensor calculations were incorporated via the ZORA.^{36–40} Shielding tensors were calculated with the gauge-independent atomic orbital (GIAO) method. All calculations were carried out with the ADF2009 package.^{41,42}

Results and Discussion

Table 1 summarizes the geometrical results obtained for the $I_h - E_{12}^{2-}$ ($E = \text{Ge}, \text{Sn}, \text{Pb}$) clusters at the BP86/TZ2P level. The largest deviation from the BP86/TZ2P level is obtained using the SO-ZORA correction for Pb_{12}^{2-} (0.08 Å). Note that while the relativistic effects induce a slightly diminution of the icosahedron inner space for the Ge_{12}^{2-} and Sn_{12}^{2-} clusters, they expand the Pb_{12}^{2-} cage. Differences between the ZORA and SO-ZORA approaches are not very significant in terms of geometrical parameters. In order to avoid any possible error source and confusion, we took the geometries optimized at the SO-ZORA-BP86/TZ2P level to calculate the shielding tensors using the different theoretical approximations.

Let us first concentrate on stannaspherene. The shielding tensor is well-defined at any position of the space. Obviously, NICS and \mathbf{B}^{ind} can also be calculated at any point of the space.^{43,44} The first one is a scalar molecular field, and the second is a vector molecular field.⁴⁵ Figures 1A and B depict the NICS and the z -component of the \mathbf{B}^{ind} (B_z^{ind}) profiles of

Table 1. Relevant Geometrical Parameters of $I_h - E_{12}^{2-}$ ($E = \text{Ge}, \text{Sn}, \text{Pb}$)^a



	Ge		Sn		Pb	
	E-E	r	E-E	r	E-E	r
BP86/TZ2P	2.745	2.611	3.132	2.979	3.169	3.014
ZORA-BP86/TZ2P	2.736	2.602	3.130	2.976	3.250	3.090
SO-ZORA-BP86/TZ2P	2.735	2.601	3.129	2.975	3.251	3.091

^a r is the distance between the cage center and one E atom. Both distances are given in Å.

Sn_{12}^{2-} , respectively. Given the spherical symmetry of this compound, the NICS and B_z^{ind} values calculated at the cage center, NICS(0) and $B_z^{\text{ind}}(0)$, using the BP86/TZ2P level are both 8.1 ppm. One can assume that the system is antiaromatic (using these magnetic descriptors), the conclusion previously reached by Chen et al.⁵ However, the relativistic effects modify strongly both electron delocalization descriptors. Using ZORA and SO-ZORA approaches, the NICS(0) and $B_z^{\text{ind}}(0)$ values are negative (-4 ppm), i.e., there is a drastic change of approximately 12 ppm. Of course, the chemical interpretation also changes; now the system can be considered slightly aromatic as was pointed out by Tian and co-workers.¹⁴ The same trend is obtained at the PW91 or SAOP levels. Obviously, there are some changes in the values, but it is possible to appreciate drastic changes in magnitude induced by the inclusion of relativistic effects (see Table 1). Interestingly, the profiles show a minimum at approximately 2.5 and 3 Å for NICS and B_z^{ind} , respectively, from the cage center, which coincides with the three-membered ring center, indicating a σ -delocalization increase at each icosahedron face, as is observed in other all-metal clusters (see also Table 1 in Supporting Information).^{46–51}

Next, we will compare the stannaspherene NICS profiles with those obtained for Ge_{12}^{2-} and Pb_{12}^{2-} . In the first case (Figure 2), the relativistic effects change by less than 6 ppm in the NICS and B_z^{ind} values; it is not unexpected that Ge, the lightest element among the series, is also the least affected by the inclusion of relativistic effects. However, for the case of plumbaspherene (see Figure 3), the addition of relativistic effects to NICS/ B_z^{ind} profiles may modify the conclusion about aromaticity of this cluster. In Figure 3 one can see that the difference from the nonrelativistic to the relativity-included calculation is almost 40 ppm! Just by examining the nonrelativistic results one would argue that Pb_{12}^{2-} is a nonaromatic molecule, however, when taking into account the relativistic effects, the large negative NICS/ B_z^{ind} values indicate a high degree of aromaticity. Moreover, by inspecting the B_z^{ind} profiles one can observe that plumbaspherene

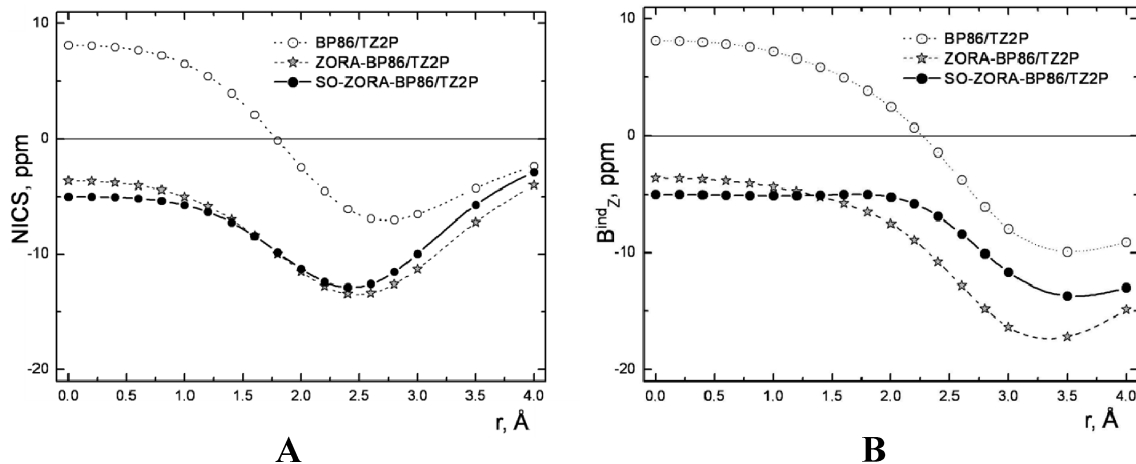


Figure 1. (A) NICS profile and (B) B_z^{ind} profile for Sn_{12}^{2-} .

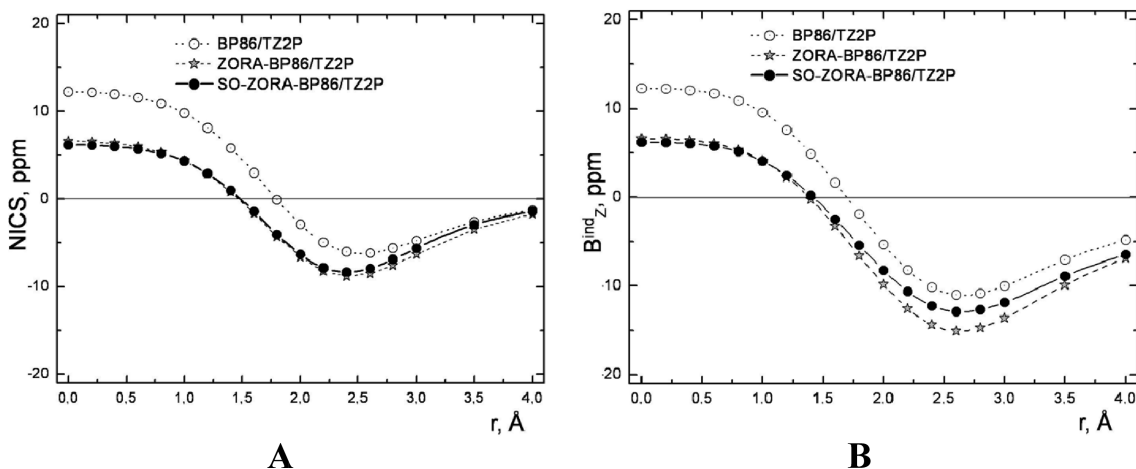


Figure 2. (A) NICS profile and (B) B_z^{ind} profile for Ge_{12}^{2-} .

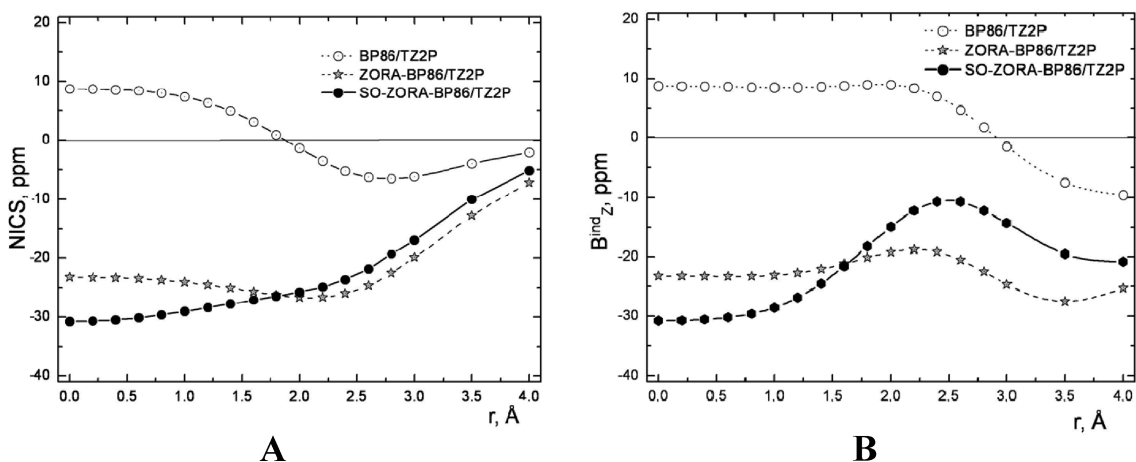


Figure 3. (A) NICS profile and (B) B_z^{ind} profile for Pb_{12}^{2-} .

possesses a maximum (between 2 and 2.5 Å) close to the center of the three-membered ring Pb–Pb–Pb (located at 3.0 Å), while its homologues show a minimum close to this region.

In several papers it is discussed whether NICS or NICS_{zz} (which is identical to B_z^{ind}) are enough to determine the aromatic character of a molecule.^{52–58} Nevertheless, these indices have become a popular probe of aromaticity in a variety of molecules. In view of the fact that the NICS and

B_z^{ind} indices are extensively used by computational and theoretically oriented experimental chemists, there is an important warning: the relativistic effects can drastically change the NICS and the B_z^{ind} values and, thus, the chemical interpretation. We have obtained profiles using PW91 and SAOP functionals, and the picture obtained is qualitatively the same, reaffirming the enhanced influence of relativistic effects in the calculations of NICS and B_z^{ind} values, no matter which functional is used (See the Supporting Information).

Table 2. The z-Component of the Induced Magnetic Field (B_z^{ind}) in ppm Calculated at the Cage Center^a

	Ge	Sn	Pb
BP86/TZ2P	12.5	8.1	8.7
ZORA-BP86/TZ2P	6.8	-3.6	-23.2
SO-ZORA-BP86/TZ2P	6.4	-5.0	-30.8
PW91/TZ2P	12.7	8.9	9.3
ZORA-PW91/TZ2P	7.0	-3.2	-23.4
SO-ZORA-PW91/TZ2P	6.6	-4.5	-30.8
SAOP/TZ2P	7.8	4.8	5.0
ZORA-SAOP/TZ2P	2.5	-6.0	-22.8
SO-ZORA-SAOP/TZ2P	2.2	-7.2	-30.9

^a Given the spherical symmetry of the title compounds, the B_z^{ind} and the NICS values calculated at the cage center are the same.

Thus, we can safely conclude that the results found are not an artifact of BP86 functional.

With this scenario one could anticipate that relativistic effects are important in any molecule containing a heavy element. Indeed, our results show that the changes observed in the NICS and B_z^{ind} values in the title cages are a consequence of the treatment of the relativistic effects. If these effects are included as scalar or spin-orbit calculations, then we can establish the following trend: Ge_{12}^{2-} is a nonaromatic species, Sn_{12}^{2-} is a low aromatic species, and Pb_{12}^{2-} is strongly aromatic. So, any prediction of electron delocalization for these clusters without considering relativistic effects is likely to be erroneous.

Acknowledgment. The work in Guanajuato was funded by DAIP-UGTO, Concyteg, and Conacyt (Grant 57892). A.C.C. thanks to CONCYTEG for the Ph.D. fellowship. Part of this work has been supported by Fondecyt (Grant 11090431) and MECESUP (Grant FSM0605). E.M. acknowledges financial support from the Marie Curie Intra-European Fellowship, Seventh Framework Programme (FP7/2007-2013), under grant agreement no. PIEF-GA-2008-221734 and from the Polish Ministry of Science and Higher Education (Project no. N N204 215634).

Supporting Information Available: The NICS and the B_z^{ind} values calculated at the face center and the NICS and Bindz profiles calculated at the PW91 and SAOP levels. This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- Cui, L. F.; Wang, L. S. *Int. Rev. Phys. Chem.* **2008**, *27*, 139.
- Cui, L. F.; Huang, X.; Wang, L. M.; Li, J.; Wang, L. S. *J. Phys. Chem. A* **2006**, *110*, 10169.
- Cui, L. F.; Huang, X.; Wang, L. M.; Zubarev, D. Y.; Boldyrev, A. I.; Li, J.; Wang, L. S. *J. Am. Chem. Soc.* **2006**, *128*, 8390.
- Chen, X.; Deng, K. M.; Liu, Y. Z.; Tang, C. M.; Yuan, Y. B.; Hu, F. L.; Wu, H. P.; Huang, D. C.; Tan, W. S.; Wang, X. *Chem. Phys. Lett.* **2008**, *462*, 275.
- Chen, Z. F.; Neukermans, S.; Wang, X.; Janssens, E.; Zhou, Z.; Silverans, R. E.; King, R. B.; Schleyer, P. v. R.; Lievens, P. *J. Am. Chem. Soc.* **2006**, *128*, 12829.
- Cui, L. F.; Huang, X.; Wang, L. M.; Li, J.; Wang, L. S. *Angew. Chem.-Int. Ed.* **2007**, *46*, 742.
- Dognon, J. P.; Clavaguera, C.; Pyykkö, P. *Angew. Chem.-Int. Ed.* **2007**, *46*, 1427.
- Kocak, F. S.; Zavalij, P.; Lam, Y. F.; Eichhorn, B. W. *Inorg. Chem.* **2008**, *47*, 3515.
- Matxain, J. M.; Piris, M.; Formoso, E.; Mercero, J. M.; Lopez, X.; Ugalde, J. M. *ChemPhysChem* **2007**, *8*, 2096.
- Sun, Z. M.; Xiao, H.; Li, J.; Wang, L. S. *J. Am. Chem. Soc.* **2007**, *129*, 9560.
- Wang, J. Q.; Stegmaier, S.; Wahl, B.; Fassler, T. F. *Chem.-Eur. J.* **2010**, *16*, 1793.
- Zdetsis, A. D. *J. Chem. Phys.* **2009**, *131*.
- Neukermans, S.; Janssens, E.; Chen, Z. F.; Silverans, R. E.; Schleyer, P. v. R.; Lievens, P. *Phys. Rev. Lett.* **2004**, *92*.
- Chen, D. L.; Tian, W. Q.; Feng, J. K.; Sun, C. C. *J. Phys. Chem. A* **2007**, *111*, 8277.
- Malkin, V. G.; Malkina, O. L.; Salahub, D. R. *Chem. Phys. Lett.* **1996**, *261*, 335.
- Schreckenbach, G.; Ziegler, T. *Int. J. Quantum Chem.* **1997**, *61*, 899.
- Kaupp, M.; Malkina, O. L.; Malkin, V. G.; Pyykkö, P. *Chem.-Eur. J.* **1998**, *4*, 118.
- Wolff, S. K.; Ziegler, T. *J. Chem. Phys.* **1998**, *109*, 895.
- Buhl, M.; Kaupp, M.; Malkina, O. L.; Malkin, V. G. *J. Comput. Chem.* **1999**, *20*, 91.
- Visscher, L.; Enevoldsen, T.; Saue, T.; Jensen, H. J. A.; Oddershede, J. *J. Comput. Chem.* **1999**, *20*, 1262.
- Wolff, S. K.; Ziegler, T.; van Lenthe, E.; Baerends, E. J. *J. Chem. Phys.* **1999**, *110*, 7689.
- Vankova, N.; Heine, T.; Kortz, U. *Eur. J. Inorg. Chem.* **2009**, 5102.
- Corminboeuf, C. *Chem. Phys. Lett.* **2006**, *418*, 437.
- Alvarado-Soto, L.; Ramirez-Tagle, R.; Arratia-Perez, R. *Chem. Phys. Lett.* **2008**, *467*, 94.
- Tsipis, A. C.; Kefalidis, C. E.; Tsipis, C. A. *J. Am. Chem. Soc.* **2008**, *130*, 9144.
- Munoz-Castro, A.; Arratia-Perez, R. *J. Phys. Chem. A* **2010**, *114*, 5217.
- Schleyer, P. v. R.; Maerker, C.; Dransfeld, A.; Jiao, H. J.; Hommes, N. J. R. v. E. *J. Am. Chem. Soc.* **1996**, *118*, 6317.
- Chen, Z. F.; Wannere, C. S.; Corminboeuf, C.; Puchta, R.; Schleyer, P. v. R. *Chem. Rev.* **2005**, *105*, 3842.
- Merino, G.; Heine, T.; Seifert, G. *Chem.-Eur. J.* **2004**, *10*, 4367.
- Heine, T.; Islas, R.; Merino, G. *J. Comput. Chem.* **2007**, *28*, 302.
- Becke, A. D. *Phys. Rev. A: At., Mol., Opt. Phys.* **1988**, *38*, 3098.
- Perdew, J. P. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1986**, *33*, 8822.
- Perdew, J. P.; Chevary, J. A.; Vosko, S. H.; Jackson, K. A.; Pederson, M. R.; Sing, D. J.; Fiolhais, C. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1992**, *46*, 6671.
- Gritsenko, O. V.; Schipper, P. R. T.; Baerends, E. J. *Chem. Phys. Lett.* **1999**, *302*, 199.

- (35) Schipper, P. R. T.; Gritsenko, O. V.; van Gisbergen, S. J. A.; Baerends, E. J. *J. Chem. Phys.* **2000**, *112*, 1344.
- (36) Chang, C.; Pelissier, M.; Durand, P. *Phys. Scr.* **1986**, *34*, 394.
- (37) Lindroth, E.; Heully, J. L.; Lindgren, I.; Martensson-Pendrill, A. M. *J. Phys. B: At., Mol. Opt. Phys.* **1987**, *20*, 1679.
- (38) van Lenthe, E.; Baerends, E. J.; Snijders, J. G. *J. Chem. Phys.* **1993**, *99*, 4597.
- (39) van Lenthe, E.; Snijders, J. G.; Baerends, E. J. *J. Chem. Phys.* **1996**, *105*, 6505.
- (40) van Lenthe, E.; van Leeuwen, R.; Baerends, E. J.; Snijders, J. G. *Int. J. Quantum Chem.* **1996**, *57*, 281.
- (41) Baerends, E. J.; Autschbach, J.; Berger, J. A.; Bérces, A.; Bickelhaupt, F. M.; Bo, C.; Boeij, P. L. D.; Boerrigter, P. M.; Cavallo, L.; Chong, D. P.; Deng, L.; Dickson, R. M.; Ellis, D. E.; Faassen, M. v.; Fan, L.; Fischer, T. H.; Guerra, C. F.; Gisbergen, S. J. A. v.; Götz, A. W.; Groeneveld, J. A.; Gritsenko, O. V.; Grüning, M.; Harris, F. E.; Hoek, P. v. d.; Jacob, C. R.; Jacobsen, H.; Jensen, L.; Kadantsev, E. S.; Kessel, G. v.; Klooster, R.; Kootstra, F.; Krykunov, M. V.; Lenthe, E. v.; Louwen, J. N.; McCormack, D. A.; Michalak, A.; Neugebauer, J.; Nicu, V. P.; Osinga, V. P.; Patchkovskii, S.; Philipsen, P. H. T.; Post, D.; Pye, C. C.; Ravenek, W.; Rodriguez, J. I.; Romaniello, P.; Ros, P.; Schipper, P. R. T.; Schreckenbach, G.; Snijders, J. G.; Solà, M.; Swart, M.; Swerhone, D.; Velde, G. t.; Vernooijs, P.; Versluis, L. V. L.; Visser, O.; Wang, F.; Wesolowski, T. A.; Wezenbeek, E. M. v.; Wiesenekker, G.; Wolff, S. K.; Woo, T. K.; Yakovlev, A. L.; Ziegler, T. *ADF 2008.01*, SCM, Theoretical Chemistry; Scientific Computing and Modelling NV, Vrije Universiteit: Amsterdam, The Netherlands, 2008.
- (42) te Velde, G.; Bickelhaupt, F. M.; Baerends, E. J.; Fonseca Guerra, C.; van Gisbergen, S. J. A.; Snijders, J. G.; Ziegler, T. *J. Comput. Chem.* **2001**, *22*, 931.
- (43) Jimenez-Halla, J. O. C.; Matito, E.; Robles, J.; Sola, M. *J. Organomet. Chem.* **2006**, *691*, 4359.
- (44) Tiznado, W.; Perez-Peralta, N.; Islas, R.; Toro-Labbe, A.; Ugalde, J. M.; Merino, G. *J. Am. Chem. Soc.* **2009**, *131*, 9426.
- (45) Merino, G.; Vela, A.; Heine, T. *Chem. Rev.* **2005**, *105*, 3812.
- (46) Boldyrev, A. I.; Wang, L. S. *Chem. Rev.* **2005**, *105*, 3716.
- (47) Chen, Z. F.; Corminboeuf, C.; Heine, T.; Bohmann, J.; Schleyer, P. v. R. *J. Am. Chem. Soc.* **2003**, *125*, 13930.
- (48) Kuznetsov, A. E.; Birch, K. A.; Boldyrev, A. I.; Li, X.; Zhai, H. J.; Wang, L. S. *Science* **2003**, *300*, 622.
- (49) Li, X.; Kuznetsov, A. E.; Zhang, H. F.; Boldyrev, A. I.; Wang, L. S. *Science* **2001**, *291*, 859.
- (50) Islas, R.; Heine, T.; Ito, K.; Schleyer, P. v. R.; Merino, G. *J. Am. Chem. Soc.* **2007**, *129*, 14767.
- (51) Islas, R.; Heine, T.; Merino, G. *J. Chem. Theory Comput.* **2007**, *3*, 775.
- (52) Lazzeretti, P. *Phys. Chem. Chem. Phys.* **2004**, *6*, 217.
- (53) Viglione, R. G.; Zanasi, R.; Lazzeretti, P. *Org. Lett.* **2004**, *6*, 2265.
- (54) Faglioni, F.; Ligabue, A.; Pelloni, S.; Soncini, A.; Viglione, R. G.; Ferraro, M. B.; Zanasi, R.; Lazzeretti, P. *Org. Lett.* **2005**, *7*, 3457.
- (55) Osuna, S.; Poater, J.; Bofill, J. M.; Alemany, P.; Sola, M. *Chem. Phys. Lett.* **2006**, *428*, 191.
- (56) Feixas, F.; Jimenez-Halla, J. O. C.; Matito, E.; Poater, J.; Sola, M. *Pol. J. Chem.* **2007**, *81*, 783.
- (57) Feixas, F.; Matito, E.; Poater, J.; Sola, M. *J. Phys. Chem. A* **2007**, *111*, 4513.
- (58) Islas, R.; Martinez-Guajardo, G.; Jimenez-Halla, J. O. C.; Sola, M.; Merino, G. *J. Chem. Theory Comput.* **2010**, *6*, 1131.

Solving the Independent-Particle Model via Nonunitary Transformation Based on Variational Coupled Cluster Singles

Jozef Noga^{*,†,‡} and Ján Šimunek[†]

Department of Inorganic Chemistry, Faculty of Natural Sciences, Comenius University, Mlynská dolina CH2, SK-84215 Bratislava, Slovakia, and Institute of Inorganic Chemistry, Slovak Academy of Sciences, SK-84536 Bratislava, Slovakia

Received June 8, 2010

Abstract: We propose an alternative new approach to obtain the Slater determinant ground state solution within an independent-particle approximation using the exponential ansatz for the wave function (Thouless theorem) and exact treatment in terms of variational coupled cluster singles. Although the resulting nonlinear equations formally represent nonterminating expansions, these can be reformulated to finite expansions in terms of the density matrix correction. The latter can be exactly calculated using a very simple recurrence relation within the occupied-occupied block, while the complementary occupied–virtual and virtual–virtual blocks are related and trivially obtained by subsequent matrix multiplications involving the amplitudes of the single-excitation operator. The density matrix is naturally idempotent in any step of the iterative procedure. Blocks of the density matrix are without any further change, apart from the sign, used in the orbital transformation matrix. The latter is not a unitary one, hence leading to nonorthogonal and unnormalized molecular orbitals. These are, however, biorthogonal and can be easily orthonormalized per blocks, if needed in the post-SCF calculations. Formulation is diagonalization free, and the implementation can be easily parallelized. Finally, the formulation provides a challenging way to the solution with “a priori” localized orbitals, a way toward a linear scaling algorithm.

1. Introduction

Recently, we have published a study on the one-particle basis set relaxation effect in the explicitly correlated coupled cluster theory.¹ Our primary goal was to show the error introduced by the assumption of the generalized Brillouin theorem when one uses the explicitly correlated R12-based methods with relatively small main computational (atomic orbital) basis sets. Among others, we have investigated the performance of the traditional coupled cluster singles (CCS) model if one starts from the reference determinant corresponding to a Hartree–Fock (HF) solution with very small (or minimal) basis set, while

in the subsequent CCS calculation the virtual space is created using a much larger basis set. Due to the Thouless theorem,² the result should be close to the HF solution with this large basis; nevertheless, it deviates from the correct solution due to the nonvariational nature of the traditional CCS solution. Our observation was that the energies were generally overestimated. Obviously, variational treatment of coupled cluster singles (VCCS) leads to the Hartree–Fock solution, but such an approach gives rise to an infinite expansion of connected terms of the effective Hamiltonian.³ In an effort to proceed along the truncated expansion in terms of perturbation theory in a generalized sense,^{4,5} we discovered that the structure of this expansion enables an effective and exact reformulation in terms of the density correction matrix that can be obtained via simple recurrence relation.⁶ Thus, unexpectedly, we arrived at an alternative diagonalization-

* To whom correspondence should be addressed. E-mail: jozef.noga@savba.sk.

[†] Comenius University.

[‡] Slovak Academy of Sciences.

free solution of the HF equations, more generally applicable for any other independent-electron self-consistent-field (SCF) model.

The procedure to solve the SCF model essentially consists of two main steps in each iteration.^{7,8} The first step is related to the construction of the effective (density-dependent) Hamiltonian (Fock/KS matrix). This step is not affected by our method. The effort directed toward linear scaling approaches resulted in several techniques applied in different contributions to the Fock/KS matrix.^{9–20} In the second step, the density matrix is updated. Conventionally, this matrix is obtained from updated occupied orbitals resulting from the diagonalization of the current Fock/KS matrix (Roothaan step²¹). For linearly scaling methods and/or parallelization, the diagonalization is an unwanted step. Two main alternatives to effectively solve this problem have been focused to direct update of the density matrix^{7,19,20,22–31} and to directly obtain the localized orbitals.^{32–35}

Our method provides an unconventional solution that is based on a nonunitary transformation of the Fock matrix. This solution is surprisingly simple. In section 2, we outline this theory followed by some demonstration of its performance in section 3 and final remarks in section 4.

2. Theory

Let us start from a normalized single reference Slater determinant (Φ) that is related to a basis set of occupied (index i, j, \dots) and virtual (index a, b, \dots) spin orbitals. For arbitrary orbitals, we shall use indices p, q, \dots . At the moment, let us assume that this basis is orthogonal. Let

$$\hat{T}_1 = \sum_{i,a} t_a^i a_i^a \quad (1)$$

be a single-excitation operator with t_a^i being the amplitudes of the particular determinant created by the action of a_i^a on $|\Phi\rangle$. Thouless theorem² says that a transformation of the Slater determinant Φ via the action of an exponential operator $e^{\hat{T}_1}$ is again a single Slater determinant. Hence, solution Ψ of any independent-particle model (SCF theory) can be obtained from Φ by finding the amplitudes of \hat{T}_1 optimal to the pertinent model.

$$|\Psi\rangle = e^{\hat{T}_1} |\Phi\rangle \quad (2)$$

This is a long known fact; however, since such ansatz gives rise to a nonunitary transformation, attention has been turned to exponential parametrization leading to unitary transformation, i.e., when the operator in the exponential is antisymmetric.⁷ Such a parametrization was successfully applied long ago, e.g., in modern MCSCF and CASSCF algorithms.^{36–38}

Direct straightforward use of eq 2 in exact solution within the Hartree–Fock model has been reported in our preliminary conference contribution.⁶ This was based on the variational coupled cluster treatment, which has led to very simple final equations, which naturally avoid the diagonalization step. Here, we recapitulate the derivation in a different perspective.

2.1. One Particle Density Matrix. We shall show later that the density matrix plays the key role in the exact solution using ansatz eq 2. Its parts constitute the building blocks for all the (nonunitary) transformation matrices leading to the exact solution for the pertinent model. Matrix elements of the one particle density matrix (\mathbf{D}) expressed in the reference basis and using the ansatz of eq 2 are given as

$$D_p^q = \frac{\langle \Phi | e^{\hat{T}_1} a_p^q e^{\hat{T}_1} | \Phi \rangle}{\langle \Phi | e^{\hat{T}_1} e^{\hat{T}_1} | \Phi \rangle} \quad (3)$$

As it was shown over 40 years ago by Čížek,³ this expectation value of the replacement operator a_p^q can be expressed in an infinite expansion of connected terms:

$$D_p^q = \langle \Phi | a_p^q | \Phi \rangle \delta_i^j + \langle \Phi | (e^{\hat{T}_1} \tilde{a}_p^q e^{\hat{T}_1})_c | \Phi \rangle \quad (4)$$

where \tilde{a}_p^q is a normal ordered replacement operator with respect to Φ . The first term represents elements of the reference density matrix (\mathbf{D}^{ref}), which is in our basis the unity matrix in the occupied–occupied (oo) block. The second term represents correction to this reference density. For nonzero \hat{T}_1 also the virtual–virtual (vv) and occupied–virtual (ov) blocks of \mathbf{D} are nonzero. Let the density correction matrix be \mathbf{X} . Then

$$\mathbf{D} = \begin{pmatrix} \mathbf{1}_{\text{oo}} + \mathbf{X}_{\text{oo}} & \mathbf{X}_{\text{ov}} \\ \mathbf{X}_{\text{vo}} & \mathbf{X}_{\text{vv}} \end{pmatrix} \quad (5)$$

2.1.1. Relations between the Density Matrix Sub-Blocks. In a spin–orbital basis matrix of eq 5 is idempotent. $\mathbf{D} = \mathbf{D}\mathbf{D}$ implies that the individual blocks are related as follows:

$$\mathbf{X}_{\text{oo}} = -\mathbf{X}_{\text{oo}}\mathbf{X}_{\text{oo}} - \mathbf{X}_{\text{ov}}\mathbf{X}_{\text{vo}} \quad (6)$$

$$\mathbf{X}_{\text{vv}} = \mathbf{X}_{\text{vv}}\mathbf{X}_{\text{vv}} + \mathbf{X}_{\text{vo}}\mathbf{X}_{\text{ov}} \quad (7)$$

$$\mathbf{X}_{\text{vv}}\mathbf{X}_{\text{vo}} = -\mathbf{X}_{\text{vo}}\mathbf{X}_{\text{oo}} \quad (8)$$

$$\mathbf{X}_{\text{ov}}\mathbf{X}_{\text{vv}} = -\mathbf{X}_{\text{oo}}\mathbf{X}_{\text{ov}} \quad (9)$$

Each matrix element of \mathbf{X} represents an infinite expansion. We naturally assume that \hat{T}_1^\dagger is a true Hermitian conjugate of \hat{T}_1 and hence for real amplitudes $t_i^a = t_a^i$. First terms in the aforementioned expansions are given by \mathbf{t} (\mathbf{t}^T) for the ov (vo) blocks, whereas in oo and vv blocks the initial terms arise from single contractions of \hat{T}_1^\dagger with \hat{T}_1 via the virtual orbital index resulting in

$$\mathbf{x}_{\text{oo}} = -\mathbf{t}\mathbf{t}^\text{T} \quad (10)$$

or via the occupied orbital index resulting in

$$\mathbf{x}_{\text{vv}} = -\mathbf{t}^\text{T}\mathbf{t} \quad (11)$$

Using diagrammatic technique it is easy to show that⁶

$$\mathbf{X}_{oo} = \sum_{n=1}^{\infty} \mathbf{x}_{oo}^n \quad (12)$$

$$\mathbf{X}_{vv} = - \sum_{n=1}^{\infty} \mathbf{x}_{vv}^n \quad (13)$$

(Note that in ref 6 we used opposite signs for the oo and vv blocks of \mathbf{X} , since the matrix has been related to the density correction operator.)

Off diagonal blocks can be easily calculated as

$$\mathbf{X}_{ov} = \mathbf{X}_{vo}^T = \mathbf{t} + \mathbf{X}_{oo}\mathbf{t} = \mathbf{D}_{oo}\mathbf{t} = \mathbf{t} - \mathbf{t}\mathbf{X}_{vv} \quad (14)$$

and vice versa the diagonal blocks are related to off-diagonal ones by simple relations:

$$\mathbf{X}_{vv} = \mathbf{X}_{vo}\mathbf{t} \quad (15)$$

$$\mathbf{X}_{oo} = -\mathbf{t}\mathbf{X}_{vo} \quad (16)$$

Identities of eqs 6–9 equivalently follow from eqs 10–14. It underlines the fact that the density matrix derived from the wave function of eq 2 is naturally idempotent without imposing this property. Moreover, it is easy to show that this matrix also fulfills the trace relation

$$\text{Tr}(\mathbf{D}) = N_{el} \quad (17)$$

If we define

$$\mathbf{y}_k = \mathbf{t}^T \mathbf{x}_{oo}^k = \mathbf{x}_{vv}^k \mathbf{t} \quad (18)$$

Equations 12 and 13 can be equivalently rewritten as

$$\mathbf{X}_{oo} = - \sum_{k=0}^{\infty} (\mathbf{y}_k \mathbf{y}_k^T + \mathbf{y}_k \mathbf{y}_{k+1}^T) \quad (19)$$

$$\mathbf{X}_{vv} = \sum_{k=0}^{\infty} (\mathbf{y}_k^T \mathbf{y}_k + \mathbf{y}_k^T \mathbf{y}_{k+1}) \quad (20)$$

Since, obviously, for any k, l

$$\text{Tr}(\mathbf{y}_k \mathbf{y}_l^T) = \text{Tr}(\mathbf{y}_l^T \mathbf{y}_k) \quad (21)$$

and consequently

$$\text{Tr}(\mathbf{X}_{oo}) = -\text{Tr}(\mathbf{X}_{vv}) \quad (22)$$

$$\text{Tr}(\mathbf{D}) = \text{Tr}(\mathbf{D}^{\text{ref}}) = N_{el} \quad (23)$$

2.1.2. Obtaining the Density Matrix. The density matrix form of eq 5 is only relevant when eqs 12 and 13 are convergent series. This is true if the eigenvalues of the \mathbf{x}_{oo} matrix are contained in the interval $(-1,0)$, and similarly, eigenvalues of the \mathbf{x}_{vv} matrix are from $(0,1)$. Consequently, the values of \mathbf{t} -amplitudes must be from $(-1,1)$. This also means that the reference determinant must dominate, since its weight in the wave function expansion is unity. Sufficiently good reference is usually achieved using a SCF start from a very small basis or simplified models. Convergence

in these expansions is quadratic in \mathbf{t} and, in addition, it is accelerated by the alternating sign of the individual contributions. Equation 14 shows that the sub-blocks of \mathbf{X} are related and, consequently, only that of \mathbf{X}_{oo} or \mathbf{X}_{vv} needs to be evaluated via the infinite expansion.

Equations 5 and 12 imply that

$$\mathbf{X}_{oo} = \mathbf{D}_{oo} \mathbf{x}_{oo} \quad (24)$$

and subsequently from eq 5

$$\mathbf{D}_{oo} = \mathbf{1}_{oo} + \mathbf{D}_{oo} \mathbf{x}_{oo} \quad (25)$$

Equation 25 suggests a recurrence relation to calculate \mathbf{D}_{oo} :

$$\mathbf{D}_{oo}^{(k+1)} = \mathbf{1}_{oo} + \mathbf{D}_{oo}^{(k)} \mathbf{x}_{oo} \quad (26)$$

starting with

$$\mathbf{D}_{oo}^{(0)} = \mathbf{1}_{oo} \quad (27)$$

Alternatively, the \mathbf{D}_{oo} block can be evaluated from eq 25 as

$$\mathbf{D}_{oo} = (\mathbf{1} - \mathbf{x}_{oo})^{-1} \quad (28)$$

Though matrix inversion is not our preferred operation, the latter evaluation of \mathbf{D}_{oo} might be useful when the recurrence of eq 26 converges very slowly. However, using eq 28 would be unphysical when eq 12 is divergent, as pointed out at the beginning of this section.

Calculation of the remaining blocks of \mathbf{D} easily follows from eqs 14 and 15. It is not the purpose of this paper to deal with the Fock matrix (\mathbf{F}) construction. Having updated \mathbf{D} and/or subsequently the electron density, this can be performed using established procedures in DFT or Hartree–Fock methods.

2.2. Orbital Transformation Matrix. For this moment, let us assume that we have the Fock matrix in the original reference molecular orbital basis and hence the Fock operator

$$\hat{F} = \sum_{p,q} F_p^q a_p^q \quad (29)$$

In fact, \hat{F} can be replaced by any generic independent-particle operator. The expectation value of this operator in the wave function of eq 2 reads

$$\frac{\langle \Phi | e^{\hat{T}_1} \hat{F} e^{\hat{T}_1} | \Phi \rangle}{\langle \Phi | e^{\hat{T}_1} e^{\hat{T}_1} | \Phi \rangle} = \sum_{p,q} F_p^q \frac{\langle \Phi | e^{\hat{T}_1} a_p^q e^{\hat{T}_1} | \Phi \rangle}{\langle \Phi | e^{\hat{T}_1} e^{\hat{T}_1} | \Phi \rangle} = \text{Tr}[\mathbf{F}\mathbf{D}(\mathbf{t}^T, \mathbf{t})] \quad (30)$$

Making this expectation value stationary with respect to the amplitudes of \hat{T}_1^{\dagger} (or equivalently \hat{T}_1) gives rise to equations determining \mathbf{t} . In this step, \mathbf{F} can be treated as fixed and independent from \mathbf{t} . Thus, our task is reduced to solve for

$$0 = \sum_{p,q} F_p^q \frac{\partial D_p^q}{\partial t_i^a} \quad (31)$$

Differentiating the oo block of \mathbf{D} using eqs 10 and 12 gives rise to

$$\sum_{k,l}^{\text{occ}} F_l^k \frac{\partial D_k^l}{\partial t_i^a} = - \sum_{k,l}^{\text{occ}} \sum_{n=0}^{\infty} \sum_{m=0}^n (\mathbf{x}_{\text{oo}}^{n-m})_k^i F_l^k (\mathbf{x}_{\text{oo}}^m)_a^l \quad (32)$$

This expression can be rearranged in a way that one clearly identifies blocks of the density matrix. Afterward, the rhs of eq 32 can be rewritten in a more compact matrix form using these blocks as

$$-(\mathbf{1}_{\text{oo}} + \sum_n \mathbf{x}_{\text{oo}}^n) \mathbf{F}_{\text{oo}} (\mathbf{1}_{\text{oo}} + \sum_m \mathbf{x}_{\text{oo}}^m) \mathbf{t} = -(\mathbf{1}_{\text{oo}} + \mathbf{X}_{\text{oo}}) \mathbf{F}_{\text{oo}} \mathbf{X}_{\text{ov}} = -\mathbf{D}_{\text{oo}} \mathbf{F}_{\text{oo}} \mathbf{D}_{\text{ov}} \quad (33)$$

Similarly, for the remaining three blocks of the density using eqs 10–14 we arrive at

$$\sum_k^{\text{occ}} \sum_c^{\text{vir}} F_c^k \frac{\partial D_c^k}{\partial t_i^a} = \sum_k^{\text{occ}} \sum_c^{\text{vir}} \sum_{n=0}^{\infty} \sum_{m=0}^n (\mathbf{x}_{\text{oo}}^m)_k^i F_c^k (\mathbf{x}_{\text{vv}}^{n-m})_a^c \quad (34)$$

$$\sum_k^{\text{occ}} \sum_c^{\text{vir}} F_c^k \frac{\partial D_c^k}{\partial t_i^a} = - \sum_k^{\text{occ}} \sum_c^{\text{vir}} \sum_{n=0}^{\infty} \sum_{m=0}^n (\mathbf{t}_{\text{vv}}^{n-m})_c^i F_c^k (\mathbf{x}_{\text{oo}}^m)_a^k \quad (35)$$

$$\sum_{c,d}^{\text{vir}} F_c^d \frac{\partial D_d^c}{\partial t_i^a} = \sum_{c,d}^{\text{vir}} \sum_{n=0}^{\infty} \sum_{m=0}^n (\mathbf{t}_{\text{vv}}^{n-m})_d^i F_c^d (\mathbf{x}_{\text{vv}}^m)_a^c \quad (36)$$

From eqs 32–36 follows the matrix form of eq 31:

$$(\mathbf{1}_{\text{oo}} + \mathbf{X}_{\text{oo}}) [\mathbf{F}_{\text{ov}} (\mathbf{1}_{\text{vv}} - \mathbf{X}_{\text{vv}}) - \mathbf{F}_{\text{oo}} \mathbf{X}_{\text{ov}}] + \mathbf{X}_{\text{ov}} [\mathbf{F}_{\text{vv}} (\mathbf{1}_{\text{vv}} - \mathbf{X}_{\text{vv}}) - \mathbf{F}_{\text{vo}} \mathbf{X}_{\text{ov}}] = \tilde{\mathbf{F}}_{\text{ov}} = 0 \quad (37)$$

This equation can be also written as

$$(\mathbf{D}\mathbf{F} - \mathbf{D}\mathbf{F}\mathbf{D})_{\text{ov}} = (\mathbf{D}[\mathbf{D}, \mathbf{F}])_{\text{ov}} = 0 \quad (38)$$

and is equivalent to the Brillouin theorem.

Equation 37 represents a part of a more general transformation of the (actual) Fock operator expressed in the reference basis to a basis in which ov (vo) block is zero, in other words, to a basis in which the Brillouin theorem holds (at least with the current fixed \mathbf{F}):

$$\tilde{\mathbf{F}} = \mathbf{Q}^T \mathbf{F} \mathbf{Q} \quad (39)$$

The transformation matrix (\mathbf{Q}) can be easily extracted from eq 37:

$$\mathbf{Q} = \mathbf{1} + \begin{pmatrix} \mathbf{X}_{\text{oo}} & -\mathbf{X}_{\text{ov}} \\ \mathbf{X}_{\text{vo}} & -\mathbf{X}_{\text{vv}} \end{pmatrix} \quad (40)$$

Transformation as in eq 39 is not a unitary one. Taking into account eqs 6–9 one can easily show that

$$\mathbf{Q}^T \mathbf{Q} = \mathbf{1} + \begin{pmatrix} \mathbf{X}_{\text{oo}} & 0 \\ 0 & -\mathbf{X}_{\text{vv}} \end{pmatrix} \quad (41)$$

which is different from the unity matrix for $\mathbf{X}_{\text{oo}} \neq 0$. Moreover, $\text{Tr}(\mathbf{Q}^T \mathbf{Q}) = N + 2\text{Tr}(\mathbf{X}_{\text{oo}})$ (cf. eq 22) is also different from the number of orbitals. This means that the molecular orbitals (\mathbf{C}) that are obtained from the reference basis (\mathbf{C}^{ref}) as

$$\tilde{\mathbf{C}} = \mathbf{C}^{\text{ref}} \mathbf{Q} \quad (42)$$

are not orthogonal within the individual occupied and virtual blocks. Moreover, these orbitals are not normalized even when the reference orbitals were normalized. The occupied and virtual blocks are biorthogonal; i.e., when \mathbf{S} is the overlap matrix in the initial generally nonorthogonal computational basis, then

$$\tilde{\mathbf{C}}_o^T \mathbf{S} \tilde{\mathbf{C}}_v = 0 \quad (43)$$

Let us stress, however, that the “new” molecular orbitals (eq 42) need not be ever constructed. Their construction is fully optional for the post-SCF purpose. Using eq 41 one can immediately construct a unitary transformation matrix as

$$\mathbf{U} = \begin{pmatrix} (\mathbf{1}_{\text{oo}} + \mathbf{X}_{\text{oo}})^{1/2} & -\mathbf{X}_{\text{ov}} (\mathbf{1}_{\text{vv}} - \mathbf{X}_{\text{vv}})^{-1/2} \\ \mathbf{X}_{\text{vo}} (\mathbf{1}_{\text{oo}} + \mathbf{X}_{\text{oo}})^{-1/2} & (\mathbf{1}_{\text{vv}} - \mathbf{X}_{\text{vv}})^{1/2} \end{pmatrix} \quad (44)$$

and hence the final orthonormal set (\mathbf{C}) of molecular orbitals:

$$\mathbf{C}_o = \mathbf{C}^{\text{ref}} (\mathbf{Q}_{\text{oo}} + \mathbf{Q}_{\text{vo}}) \mathbf{Q}_{\text{oo}}^{-1/2} \quad (45)$$

$$\mathbf{C}_v = \mathbf{C}^{\text{ref}} (\mathbf{Q}_{\text{vv}} + \mathbf{Q}_{\text{ov}}) \mathbf{Q}_{\text{vv}}^{-1/2} \quad (46)$$

Note that in order to solve the problem, we as well never need the whole Fock matrix in the “updated orbital basis”, merely its ov block $\tilde{\mathbf{F}}_{\text{ov}}$ (cf. eq 37), which is obtained in a two-step matrix multiplication formally scaling as $N_o N_v^2 + N_o N_v N$, where N , N_o , and N_v are the number of basis functions, the number of occupied, and the number of virtual orbitals, respectively.

Let us recall that the energy is calculated using the correctly normalized density matrix of eq 5 as an expectation value of the Hamiltonian.

2.3. Updating t-Amplitudes. In principle, there is no explicit need to update the amplitudes, since one can get the solution for \mathbf{X} (and hence update the density) using the set of nonlinear equations resulting from eqs 37, 6, and 7. A much simpler approach, and algorithmically much more favorable, is to evaluate \mathbf{t} and subsequently the \mathbf{X}_{oo} matrix using eqs 10 and 25 from which the \mathbf{X}_{ov} and \mathbf{X}_{vv} blocks are easily obtained via eqs 14 and 15.

Indeed, in terms of \mathbf{t} , eq 37 can be rewritten as

$$\mathbf{f}_{\text{ov}} + \mathbf{t}_{\text{vv}} - \mathbf{f}_{\text{oo}} \mathbf{t} + \text{higher order terms in } \mathbf{t} = 0 \quad (47)$$

We denote by \mathbf{f} the original Fock matrix in the reference basis constructed by using the pertinent reference density. Moving the terms with diagonal elements of \mathbf{f} in eq 47 to the rhs suggests the \mathbf{t} in the $(k+1)$ iteration as

$$t_a^{i(k+1)} = t_a^{i(k)} + \tilde{F}_a^{i(k)} / (f_i^i - f_a^a) \quad (48)$$

With the initial $t_a^{i(0)} = 0$ we have

$$t_a^{i(1)} = f_a^i / (f_i^i - f_a^a) \quad (49)$$

Both the density from which the Fock matrix was calculated and the transformation matrix \mathbf{Q} are functions of the same $\mathbf{t}^{(k)}$; i.e. $\tilde{\mathbf{F}}^{(k)} \equiv \tilde{\mathbf{F}}(\mathbf{t}^{(k)})$. Hence, a one-step update

due to eq 48 is fully appropriate and leads to a balanced treatment such as that used in established coupled cluster algorithms.

Alternatively, one could solve eq 48 in a subiterative procedure with $\mathbf{F}^{(k)}$ being fixed, while merely updating the $\tilde{\mathbf{F}}_{\text{ov}}$ via eq 39 until the convergence of \mathbf{t} in the given macroiteration. We have experimented in this way. However, it turned out that the convergency of the global iterative process worsened, which suggests that such a procedure not be recommended.

2.4. Algorithm Summary. The key new aspects of this theory are (i) a facile recurrence relation of eq 35 to update the one particle density, (ii) a simple update of the \mathbf{t} -amplitudes via eq 48, which replaces the diagonalization step, and (iii) nonunitary transformation of eq 39. The main attractive feature is that the transformation matrix \mathbf{Q} (eq 40) is easily constructed from parts of the density matrix \mathbf{D} (eq 5) without additional (relevant) computational demand. These features give rise to an efficient and simple algorithm that can be implemented in any existing code by simple modifications:

Init ($k = 0$). Create the reference (initial) orbitals and the initial Fock matrix ($\mathbf{F}^{(0)} = \mathbf{f}$). It is not necessary from the principle, but it improves the convergency when one starts from semicanonical orbitals.³⁹

Step 1 ($k = k + 1$). Calculate $\mathbf{t}^{(k)}$ according to eq 48.

Step 1a. When appropriate, apply DIIS⁴⁰ to \mathbf{t} amplitudes.

Step 2. Calculate \mathbf{x}_{oo} (eq 10) and \mathbf{D}_{oo} using eq 25, \mathbf{X}_{ov} using eq 14, and \mathbf{X}_{vv} using eq 15. Complete the density matrix $\mathbf{D}^{(k)}$ (eq 5) and the transformation matrix $\mathbf{Q}^{(k)}$ (eq 40).

Step 3. Calculate $\mathbf{F}^{(k)} \equiv \mathbf{F}^{(k)}(\mathbf{D}^{(k)})$ related to the chosen independent particle model. In most algorithms, this step is performed in AO basis and a transformation of the density matrix to the latter basis precedes this step. Optionally, at this step calculate the energy and go to the final step if converged.

Step 4. Create $\tilde{\mathbf{F}}_{\text{ov}}^{(k)}$ using (partial) transformation of eq 39. If $\|\tilde{\mathbf{F}}_{\text{ov}}^{(k)}\| > \text{threshold}$, go to step 1.

Final. Calculate the energy from the resulting density and if necessary construct the molecular orbitals using eqs 45 and 46.

2.5. Discussion. Our approach does not affect the calculation of the Fock matrix. It provides a simple alternative to existing methods that replace the “traditional” diagonalization step of the Fock matrix in which the molecular orbitals are updated. Certainly, several common features can be found in these methods. Indeed, quite a similar central equation as our eq 37 appears also in exact reformulation of the diagonalization step using Cayley-type parametrization of the unitary matrix by Liang and Head-Gordon.²⁸ As follows from the preceding section, our approach also provides an exact reformulation of this problem. Consequently, the updated density matrix naturally obeys the trace relation and is idempotent without subsequent purification. The latter is needed in approaches that use a truncated parametrical expansion of the density matrix such as that used in the recent augmented Roothan-Hall method.³¹

We deviate from other methods from the beginning by working with a wave function that is not normalized. The exponential ansatz of the wave function is naturally related to the infinite expansion of the density matrix that can be exactly calculated using the recurrence formula of eq 26.

Instead, in the method of Liang and Head-Gordon, this step requires an inversion of a matrix equally in the oo block. It would be inappropriate to compare the “theoretical” count of multiplications in the aforementioned methods, as this number varies in solving the set of nonlinear equations similar to eq 37. An alternative to solving this equation is in our method provided by a simple update of the \mathbf{t} amplitudes via dividing the ov block of the Fock matrix by orbital energy denominator. If needed for a better convergence, e.g., when the (reference) HOMO is very close to LUMO, there is a space here to apply a denominator shift.

Eventually, used exponential ansatz can be effectively combined with a Newton method, since the exact second derivative of the density matrix with respect to \mathbf{t} nicely factorizes. In terms of \mathbf{Q} we have (eqs 32–36):

$$(\delta_k^q + \delta_c^q) \frac{\partial D_q^p}{\partial t_i^a} = (\delta_k^q + \delta_c^q) \frac{\partial X_q^p}{\partial t_i^a} = (\delta_k^q - \delta_c^q) Q_q^i Q_a^p \quad (50)$$

$$\frac{\partial^2 D_q^p}{\partial t_i^a \partial t_j^b} = Q_q^i Q_a^j Q_b^p + Q_q^j Q_a^i Q_b^p \quad (51)$$

3. Sample Calculations

The correctness of our approach has been numerically checked and proven already in our preliminary paper.⁶ Its performance is here demonstrated using the Hartree–Fock SCF approach for four systems including uracil, a complex of four guanine molecules, a dimer of two hydrocarbon chains ($\text{C}_{18}\text{H}_{38}$)₂, and a complex of $\text{C}_{54}\text{H}_{18}$ sheet with cytosine–guanine pair. These systems range from 12 to 112 atoms and the number of occupied orbitals varies from $N_o = 29$ for uracil to $N_o = 239$ for the last mentioned complex.

We have tested starts from various (reference) wave functions including simplified models (such as EHT or LDA approximations), small or minimal atomic orbital basis subsets, and/or a combination of both. When we started from a basis set smaller than the target computational basis set, i.e., from its subset, the virtual subspace was completed by Schmidt orthogonalizing the complementary atomic orbitals to the MO’s resulting from the small initial basis. Semicanonical orbitals³⁹ were finally created by separate diagonalization of the oo and vv block of the Fock matrix obtained using the initial density matrix. These semicanonical orbitals served as our reference basis. At this stage, we have not investigated alternative reference bases that would eventually not require a diagonalization step.⁴¹ Without any further investigation and optimization, we have used the DIIS procedure applied to the \mathbf{t} amplitudes, always from the three subsequent iterations.

Results are summarized in Table 1. Geometries are available in the Supporting Information and additional computational details are as follows:

Table 1. Performance of the Proposed Theory for Selected Molecular Systems^a

initial MOS ^b	ΔE_{ini}^c	$\Delta E_{\text{ini}} + E^{(2)}^d$	max(t)	it _{rec} ^e	it _{SCF} ^f
uracil: aug-cc-pV5Z, $N = 1336$, $N_0 = 29$, $E_{\text{HF}} = -412.655\ 211$					
EHT 2s1p/1s ($N = 44$)	2.031 453	-0.523 032	0.1327	14/22	21/26
LDA 2s1p/1s	0.942 645	0.117 791	0.1067	5/8	9/16
HF 2s1p/1s	0.890 455	0.072 809	0.0672	5/7	9/16
(C ₁₈ H ₃₈) ₂ : 6-31G**, $N = 884$, $N_0 = 146$, $E_{\text{HF}} = -1407.668\ 375$					
EHT 2s1p/1s ($N = 256$)	23.596 339	1.020 783	0.2871	10/15	7/11
LDA 2s1p/1s	20.750 178	-0.064 462	0.2770	12/18	8/11
HF 2s1p/1s	20.566 948	0.612 301	0.2937	11/17	7/11
EHT 6-31G**	6.482 139	-0.255 941	0.2066	6/10	7/11
(guanine) ₄ : 6-31G**, $N = 716$, $N_0 = 156$, $E_{\text{HF}} = -2157.142\ 890$					
EHT 6-31G ($N = 436$)	9.418 091	-0.965 297	0.1172	10/15	17/24
LDA 6-31G	0.887 659	0.144 224	0.0798	4/6	9/16
HF 6-31G	0.479 354	0.059 236	0.0361	4/6	7/11
EHT 6-31G**	9.591 119	-1.052 849	0.1189	10/16	17/24
[(C ₅₄ H ₁₈)-(cytosine-guanine)]: $N_0 = 239$					
aug-cc-pVDZ/aug-cc-pVDZ-RI, $N = 1931$, $E_{\text{HF}} = -2988.139\ 195$					
EHT 2s1p/1s ($N = 393$)	11.312 142	-0.336 064	0.1216	9/14	17/28
HF 2s1p/1s	5.761 793	0.424 057	0.0401	6/8	13/20
HF 3s2p/2s ($N = 713$)	1.300 318	0.146 213	0.0199	4/6	9/18
aug-cc-pVTZ/aug-cc-pVTZ-RI, $N = 4002$, $E_{\text{HF}} = -2988.686\ 645$					
EHT 2s1p/1s	11.571 240	-0.359 279	0.1262	9/14	17/20
HF 2s1p/1s	6.002 295	0.417 410	0.0464	6/8	13/16
HF 4s3p/3s ($N = 1033$)	1.493 402	0.172 046	0.0145	4/6	9/12

^a Energies are in E_h . ^b Approximation and basis set used to obtain the initial MOS. ^c $\Delta E_{\text{ini}} = E_{\text{ini}} - E_{\text{HF}}$; $E_{\text{ini}} = \langle \Phi | \hat{H} | \Phi \rangle$. ^d $E^{(2)}$ according to eq 52. ^e Number of recurrence cycles to converge the norm of \mathbf{D}_{oo} (eq 26) residual below $10^{-6}/10^{-10}$. ^f Number of SCF iterations a/b ; a : HF energy threshold 10^{-6} ; b : residual **t** norm threshold 10^{-5} , for aug-cc-pVTZ 10^{-4} .

Uracil: we have used the geometry of the neutral molecule as in a recent paper by Bachorz and Klopper.⁴² Here, we employed relatively extensive aug-cc-pV5Z,⁴³⁻⁴⁵ whereas the initial basis was a minimal one. 2s1p for a non-hydrogen atom and 1s for hydrogen correspond to the first contracted functions (with highest exponents) from the cc-pV5Z⁴³ set, pertinent to the given angular momentum.

(C₁₈H₃₈)₂: a complex of two parallel linear C₁₈ alkane chains⁴⁶ was calculated with the 6-31G** basis.^{47,48} The minimal 2s1p/1s basis used for constructing the initial orbital set was again the subset of the main 6-31G** basis including contracted functions with the largest exponents.

(Guanine)₄: a planar complex of four guanine molecules⁴⁹ was calculated using 6-31G** with special polarization functions with d -exponent 0.25 for non-hydrogen atoms and p -exponent 1.1 for hydrogen atoms.⁵⁰

[(C₅₄H₁₈)-(cytosine-guanine)] (a stacking van der Waals complex of a graphene sheet with a DNA base pair):⁴⁶ We show performance of our method using aug-cc-pVDZ and aug-cc-pVTZ basis sets. In these cases, the Fock matrix calculation has been performed using the RI (density fitting)

approach using the fitting sets of Weigend et al.⁵¹ Initial basis sets are subsets of the full sets, similarly as aforementioned.

In Table 1 we focus on the following quantities:

- Deviation of the energy corresponding to the initial guess E_{ini} from the exact (final) HF energy in the given basis, which is one of the measures that reflect the quality of the reference orbital basis.
- Deviation of the energy using a simple second-order energy correction from first order **t**, $E_{\text{ini}} + E^{(2)}$. This correction is given by a simple formula:

$$E^{(2)} = \sum_{i,a} (f_i^a)^2 / (f_i^i - f_a^a) \quad (52)$$

- The largest **t**-amplitude, which is a different measure of the deviation of the reference orbital basis from optimal molecular orbitals.
- The number of microiterations needed to converge the recurrence relation eq 26. Since this step includes matrix multiplications only over the oo block, the timing is typically negligible compared to the timing of the whole iteration step.
- The number of (macro)iterations using different thresholds to stop the convergence process.

One observes that initial orbitals stemming from minimal (or very small) basis sets were more than appropriate in order to achieve a smooth convergence. As the largest **t**-amplitudes demonstrate, the initial reference states were often quite far from the exact solution. The number of needed iterations both in the recurrence cycle and the outer loop loosely correlates with the error of the initial energy of the reference state and hence with the value of the largest **t**-amplitude. The EHT start seems to perform somewhat worse than HF or LDA.

As far as the recurrence iterations are concerned, one has to keep in mind that these involve matrix multiplications with the dimension of the number of occupied orbitals, i.e., the timing is practically negligible. For the least favorable case from our examples, (guanine)₄, the recurrence cycle represents less than 1% of the computational costs for matrix multiplication with the full basis. In all other examples, this fraction is still less by one or more orders of magnitude.

The column with the simple second-order energy correction is given here just for curiosity. Although it recovers most of the initial energy error in absolute value, this correction is still unreliable and, as seen, is hardly predictable mainly due to its nonvariational nature.

We add a short remark related to the convergence behavior of the suggested algorithm for basis sets exhibiting significant linear dependence. Indeed, the complex of the graphene sheet with the DNA base pair calculated with the aug-cc-pVTZ basis represents such a case and has been suggested to us as a really difficult one. The condition number for the overlap matrix is on the order of 10^{13} . Even without eliminating any functions from this ill-conditioned set, a smooth convergence has been achieved, though, as noted in Table 1, the numerical accuracy of the final **t**-amplitudes has been little bit worsened. Nevertheless, an error of 10^{-4} in the norm of **t** still guarantees that the density matrix is sufficiently converged to about 10^{-8} .

4. Conclusions

We have shown that an effective alternative way of solving the independent particle problem without diagonalization is provided by an exact solution using a variational coupled cluster singles treatment. The formal nonterminating expansion of connected terms in the resulting energy expression can be reformulated to finite expansion in terms of the density matrix correction. The latter can be exactly calculated using a very simple recurrence relation within the occupied–occupied block, while the complementary occupied–virtual and virtual–virtual blocks are related and trivially obtained by subsequent matrix multiplications involving the amplitudes of the single-excitation operator. The density matrix is naturally idempotent in any step of the iterative procedure. As well, the trace relation is satisfied at each step.

Differentiation of the energy with respect to the amplitudes of the single excitation operator leads to nonunitary transformation of the Fock matrix, however, by removing the occupied–virtual block, i.e., transforming to a basis in which the Brillouin theorem is satisfied. Such transformation leads to an unnormalized and nonorthogonal, though still biorthogonal, set of molecular orbitals. The transformation matrix is built from blocks of the density matrix without any further change, apart from the sign. If needed, final molecular orbitals are easily obtained by a simple separate orthogonalization for the occupied and virtual block.

An advantage of the present formulation is a fact that the nonlinear eq 37 does not need to be explicitly solved in each iteration, instead a simple update of the amplitudes serves well. This leads to a balanced treatment when the density and the transformation matrix are always constructed from the same amplitudes.

Since the formulation is diagonalization free, the main parallelization bottleneck is overcome. If we do not comment on the construction of the Fock matrix, the rest is fully driven by matrix multiplications and can be effectively parallelized. In the simplest case, just the parallel BLAS subroutines are used. Finally, the formulation provides a challenging way to the solution with “a priori” localized orbitals, a way toward linear scaling algorithm.

Our pilot results are very promising and confirm that implementation of the present approach into an efficient large-scale production code might be worth of considering.

Acknowledgment. This work has been supported by the Grant Agency of the Ministry of Education of the Slovak Republic and Slovak Academy of Sciences (VEGA project No. 2/0079/09 as well as by the Slovak Research and Development Agency (LPP0031-07). This work has benefitted from the Centers of Excellence program of the Slovak Academy of Sciences (COMCHEM, Contract no. II/1/2007).

Supporting Information Available: Geometries for the systems discussed. This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- (1) Noga, J.; Šimunek, J. *Chem. Phys.* **2009**, *356*, 1–6.
- (2) Thouless, D. J. *Nucl. Phys.* **1960**, *21*, 225–232.

- (3) Čížek, J. *Adv. Chem. Phys.* **1969**, *14*, 35–89.
- (4) Bartlett, R. J.; Noga, J. *Chem. Phys. Lett.* **1988**, *150*, 29–36.
- (5) Bartlett, R. J.; Kucharski, S. A.; Noga, J.; Watts, J. D.; Trucks, G. W. Some consideration of alternative ansätze in coupled-cluster theory. In *Lecture Notes in Chemistry*, 1st ed.; Kaldor, U., Ed.; Springer-Verlag: Berlin, 1989; Vol. 52, pp 125–149.
- (6) Šimunek, J.; Noga, J. *AIP Conf. Proc.* **2010**, in press.
- (7) Helgaker, T.; Jørgensen, P.; Olsen, J. Hartree-Fock Theory. In *Molecular Electronic-Structure Theory*; John Wiley: Chichester, England, 2000; Chapter 10, pp 433–513.
- (8) Parr, R. G.; Yang, W. The Kohn–Sham Method: Basic Principles. In *Density-Functional Theory of Atoms and Molecules*; Oxford University Press: New York, 1989; Chapter 7, pp 142–168.
- (9) White, C. A.; Johnson, B. G.; Gill, P. M. W.; Head-Gordon, M. *Chem. Phys. Lett.* **1996**, *253*, 268–278.
- (10) Strain, M. C.; Scuseria, G. E.; Frisch, M. J. *Science* **1996**, *271*, 51–53.
- (11) Challacombe, M.; Schwegler, E. *J. Chem. Phys.* **1997**, *106*, 5526–5536.
- (12) Shao, Y.; Head-Gordon, M. *Chem. Phys. Lett.* **2000**, *323*, 425–433.
- (13) Schwegler, E.; Challacombe, M.; Head-Gordon, M. *J. Chem. Phys.* **1997**, *106*, 9708–9717.
- (14) Schwegler, E.; Challacombe, M. *J. Chem. Phys.* **1999**, *111*, 6223–6229.
- (15) Ochsenfeld, C.; White, C. A.; Head-Gordon, M. *J. Chem. Phys.* **1998**, *109*, 1663–1669.
- (16) Burant, J. C.; Scuseria, G. E.; Frisch, M. J. *J. Chem. Phys.* **1996**, *105*, 8969–8972.
- (17) Pérez-Jordá, J. M.; Yang, W. *Chem. Phys. Lett.* **1995**, *241*, 469–476.
- (18) Stratman, R. E.; Scuseria, G. E.; Frisch, M. J. *Chem. Phys. Lett.* **1996**, *257*, 213–223.
- (19) Goedecker, S. *Rev. Mod. Phys.* **1999**, *71*, 1085–1123.
- (20) Goedecker, S.; Scuseria, G. E. *Comput. Sci. Eng.* **2003**, *5*, 14–21.
- (21) Roothaan, C. C. J. *Rev. Mod. Phys.* **1951**, *23*, 69–89.
- (22) Li, X.-P.; Nunes, R. W.; Vanderbilt, D. *Phys. Rev. B* **1993**, *47*, 10891–10894.
- (23) Millam, J. M.; Scuseria, G. E. *J. Chem. Phys.* **1997**, *106*, 5569–5577.
- (24) Scuseria, G. E. *J. Phys. Chem. A* **1999**, *103*, 4782–4790.
- (25) Challacombe, M. *J. Chem. Phys.* **1999**, *110*, 2332–2342.
- (26) Helgaker, T.; Larsen, H.; Olsen, J.; Jørgensen, P. *Chem. Phys. Lett.* **2000**, *327*, 397–403.
- (27) Liang, W. Z.; Sarvanah, C.; Shao, Y.; Baer, R.; Bell, A. T.; Head-Gordon, M. *J. Chem. Phys.* **2003**, *119*, 4117–4125.
- (28) Liang, W. Z.; Head-Gordon, M. *J. Chem. Phys.* **2004**, *120*, 10379–10384.
- (29) Köhalmi, D.; Szabados, A.; Surján, P. R. *Phys. Rev. Lett.* **2005**, *95*, 013002–4.
- (30) Sałek, P.; Thøgersen, L.; Jørgensen, P.; Manninen, P.; Olsen, J.; Jansík, B.; Reine, S.; Pawłowski, F.; Tellgren, E.; Helgaker, T.; Coriani, S. *J. Chem. Phys.* **2007**, *126*, 114110–16.

- (31) Høst, S.; Olsen, J.; Jansík, B.; Thørgensen, L.; Jørgensen, P.; Helgaker, T. *J. Chem. Phys.* **2008**, *129*, 124106–12.
- (32) Mauri, F.; Galli, G.; Car, R. *Phys. Rev. B* **1993**, *47*, 9973–9976.
- (33) Stewart, J. J. P. *Int. J. Quantum Chem.* **1996**, *58*, 133–146.
- (34) Yang, W. *Phys. Rev. B* **1997**, *56*, 9294–9297.
- (35) VandeVondele, J.; Hutter, J. *J. Chem. Phys.* **2003**, *118*, 4365–4369.
- (36) Roos, B. O.; Taylor, P. R.; Siegbahn, P. E. M. *Chem. Phys.* **1980**, *48*, 157–173.
- (37) Roos, B. O. In *Advances in Chemical Physics; Ab Initio Methods in Quantum Chemistry*; Lawley, P. K., Ed.; John Wiley: Chichester, England, 1987; Part II, pp 399.
- (38) Roos, B. O. In *Lecture Notes in Chemistry*; Roos, B. O., Ed.; Springer: Berlin, Germany, 1992; Vol. 58, pp 177.
- (39) Handy, N. C.; Pople, J. A.; Head-Gordon, M.; Raghavachari, K.; Trucks, G. W. *Chem. Phys. Lett.* **1989**, *164*, 185–192.
- (40) Pulay, P. *Chem. Phys. Lett.* **1980**, *73*, 393–398.
- (41) Szekeres, Zs.; Mezey, P. G.; Surján, P. R. *Chem. Phys. Lett.* **2006**, *424*, 420–424.
- (42) Bachorz, R. A.; Klopper, W.; Gutowski, M. *J. Chem. Phys.* **2007**, *126*, 85101–7.
- (43) Dunning, T. H., Jr. *J. Chem. Phys.* **1989**, *90*, 1007–1023.
- (44) Kendall, R. A.; Dunning, T. H., Jr.; Harrison, R. J. *J. Chem. Phys.* **1992**, *96*, 6796–6806.
- (45) Peterson, K. A.; Dunning, T. H., Jr. *J. Chem. Phys.* **2002**, *117*, 10548–10560.
- (46) Pitoňák, M. personal communication.
- (47) Hariharan, P. C.; Pople, J. A. *Theor. Chim. Acta* **1973**, *28*, 213–222.
- (48) Kroon-Batenburg, L. M. J.; van Duijneveldt, F. B. J. *J. Mol. Struct.* **1985**, *121*, 185–199.
- (49) Otero, R.; Schöck, M.; Molina, L. M.; Lægsgaard, E.; Stensgaard, I.; Hammer, B.; Besenbacher, F. *Angew. Chem., Int. Ed.* **2005**, *44*, 2270–2275.
- (50) Hobza, P.; Melhorn, A.; Cársky, P.; Zahradník, R. *J. Mol. Struct.* **1986**, *138*, 387–399.
- (51) Weigend, F.; Köhn, A.; Hättig, C. *J. Chem. Phys.* **2002**, *116*, 3175–3183.

CT1003143

Multiple Solutions to the Single-Reference CCSD Equations for NiH

Nicholas J. Mayhall and Krishnan Raghavachari*

Department of Chemistry, Indiana University, Bloomington, Indiana 47405

Received June 11, 2010

Abstract: It is typically assumed that once a Hartree–Fock (HF) reference wave function is determined, the correlated wave function obtained from that HF wave function describes the same electronic state. In this paper, we report the appearance of multiple CCSD solutions obtained from the UHF reference wave function for the known ground state of a chemically interesting molecule, NiH. To determine a correspondence between the computed CCSD solutions and the physical electronic states, we consider several characteristics of the CCSD wave functions, e.g., potential energy curves, spin density isovalue plots, and excited state studies via EOM-CCSD calculations. Finally, the use of Brueckner orbitals is encouraged as a way to avoid some of the problems highlighted here for HF-based coupled cluster calculations in such challenging systems.

1. Introduction

Coupled-cluster theory^{1–3} has undoubtedly provided electronic structure theorists with the most useful hierarchy of methods for obtaining highly accurate descriptions of electron correlation for a large variety of molecular systems. However, the nonlinearities in the wave function expansion coefficients makes it highly challenging to enumerate the multiple solutions of the resulting equations and can sometimes lead to complex behavior. For CI (configuration-interaction), which is linear in the coefficients or amplitudes, the lowest energy wave function for a particular set of molecular orbitals (MOs) can be solved for readily, as it requires the diagonalization of a Hermitian matrix. This is not the case for the CC (coupled-cluster) equations, and multiple solutions may be obtained by starting with different sets of initial amplitudes. The reasons responsible for the existence of multiple solutions in the CC case are different than those for the CI problem. The multiple solutions in CI are simply the different eigenvectors of the CI matrix. The multiple solutions to the CC equations arise from the nonlinear nature of the CC amplitude equations.

This problem was perhaps first addressed in 1978 when Monkhorst and Zivkovic⁴ explored the mathematical connections between CI and CC solutions. More recently,

Bartlett and co-workers⁵ and then Jankowski and co-workers^{6–8} studied particular examples of multiple CC solutions with the widely used H4 model⁹ system in which the geometry (comprised of four hydrogen atoms) is completely determined by a single chosen parameter. More recently, the existence of multiple CC solutions has been observed in the PPP^{10,11} model of conjugated rings.^{12–15} From these studies, it was concluded that determining a connection between a physical electronic state and a particular CC solution is a difficult problem and that a given CC solution may not even correspond to a physical state.

Multiple CC solutions also occur in a different context. In addition to obtaining solutions which differ only in the converged parameters of the multideterminantal wavefunction, one may obtain multiple CC solutions by using a different underlying HF (Hartree–Fock) reference wave function. As the HF method also requires one to solve a set of nonlinear equations, the possibility of multiple solutions arises here as well. While the choice of reference does not matter when the full n -particle T operator is used (as this is equivalent to a full configuration interaction which is invariant to rotations of all the orbitals), any realistic calculation must approximate T by a small number of excitation operators, and thus dependence on the reference wave function arises. This has been investigated by Jankowski et al.^{16–18} again using the H4 model system.

* To whom correspondence should be addressed. E-mail: kraghava@indiana.edu.

Scheme 1. Molecular Orbital Diagram for the Lowest Energy States with Δ Symmetry^a

State	(I) ${}^2\Delta$	(I) ${}^4\Delta$	(II) ${}^2\Delta$
Configuration	$d^9s^1 + 1s$	$d^8s^2 + 1s$	$d^8s^2 + 1s$
σ_{4s}	\uparrow	$\uparrow\downarrow$	$\uparrow\downarrow$
σ_{3d}	$\uparrow\downarrow$	\uparrow	\uparrow
δ_{3d}	$\uparrow\downarrow$	$\uparrow\downarrow$	$\uparrow\downarrow$
π_{3d}	$\uparrow\downarrow$	$\uparrow\downarrow$	$\uparrow\downarrow$

^a The red arrows represent electrons which participate in bonding, e.g., occupy a bonding orbital (covalent) or transfer to the H atom (ionic).

While the H4 model system is well studied and understood, there exists a dearth of information regarding the appearance of multiple CC solutions in well-studied real molecular systems. As a molecule of both chemical and physical interest, NiH has been the focus of computational and experimental investigation for many years.^{19–35} In this paper, we report the calculation of multiple solutions to the CCSD amplitude equations for the known ground state of the NiH molecule. We provide an analysis of the resulting wave functions which suggests particular implications for coupled cluster-based applications.

2. Potential Energy Surfaces

Many of the computational difficulties experienced with NiH ultimately stem from problems in correctly describing the atomic state separations of Ni.³⁶ Experimentally, the ${}^3D(d^9s^1)$ and ${}^3F(d^8s^2)$ states are nearly degenerate, with the d^9s^1 slightly more stable by 0.03 eV.³⁷ Using unrestricted HF theory (UHF) with the G3Large basis set,³³ the absence of electron correlation among the d electrons causes the d^8s^2 state to lie 1.41 eV lower in energy than the d^9s^1 state. The error in the calculated atomic state separation is reduced dramatically when correlation effects are included with CCSD(T) (0.13 eV).³³ It should be noted that, in addition to correlation effects, relativistic effects are known to contribute significantly to this energy difference (by ~ 0.3 eV).³⁸ However, as our focus in this paper is on addressing issues related to solving the CC equations, we illustrate our ideas using simple nonrelativistic calculations.

Since the ground electronic state of NiH is known to be ${}^2\Delta$, we restrict our study to the low energy states with Δ symmetry. The MO diagrams for the three lowest energy states with Δ symmetry are given in Scheme 1. On the basis of previous studies on this molecule, these electronic states can be described as follows:

1. Ground State ${}^2\Delta$. The nickel hydride molecule has a ${}^2\Delta$ ground electronic state in which both the d^9s^1 and d^8s^2 atomic states of Ni contribute to the bonding, with the d^9s^1 atomic state as the predominant component.²⁵ In a covalent bonding model, the Ni–H bond in this electronic state can be thought of principally as a Ni s orbital overlapping with a H s orbital ($4s+s$). Alternatively, in an ionic model, the Ni–H bond can be

thought of as a bond between a Ni^+ (d^9s^0) and a H^- (s^2). We will refer to this state as the ${}^2\Delta(d^9s^1)$ state.

2. Excited State ${}^2\Delta$. An excited ${}^2\Delta$ state exists which can be thought of as the d^8s^2 atomic Ni state interacting with a hydrogen atom and forming a bond between H s and Ni $d_{z^2}(d_{z^2}+s)$.³¹ However, recent large scale multireference calculations by Zou and Liu³² suggest that this state has substantial multiradical character. It may be better considered as a bond between the Ni^+ (d^8s^1) and H^- (s^2) ionic fragments resulting in a state with three unpaired electrons, although the overall spin is only 1/2. This state has been experimentally measured at 2.01 eV above the ground state.³⁹ We will refer to this state as the ${}^2\Delta(d^8s^2)$ state.
3. Excited State ${}^4\Delta$. An excited ${}^4\Delta$ state can also result from a similar atomic configuration as the ${}^2\Delta(d^8s^2)$ state above. The arrows shown in red become paired at bond lengths near the equilibrium geometry, and thus the overall multiplicity is 4. Using multireference methods, Zou and Liu calculated this state to lie 1.57 eV higher in energy than the ground state.³² This state will be denoted as ${}^4\Delta(d^8s^2)$.

2.1. UHF Solutions. Both the ground and excited ${}^2\Delta$ states can be calculated with the UHF method by starting with the appropriate orbital occupations. However, this is not without difficulty since the d_{z^2} orbital as well as the 4s orbital belong to the σ representation of the $C_{\infty v}$ point group for NiH. This leads to considerable mixing between them resulting in convergence difficulties, particularly near equilibrium. In fact, all points on the ${}^2\Delta(d^9s^1)$ potential energy surface (PES) could not be calculated with the UHF method due to convergence problems for bond lengths less than about 1.56 Å. For this region of the PES, the SCF procedure either collapsed to the lower energy ${}^2\Delta(d^8s^2)$ or simply failed to converge. However, at larger internuclear distances where the two orbitals are fairly distinct, appropriate occupation of the orbitals leads to the two different ${}^2\Delta$ states. No convergence problems arose, however, for the ${}^4\Delta(d^8s^2)$ state. In this quartet, the $\sigma_{d_{z^2}}$ and σ_{4s} orbitals are both singly occupied.

UHF potential energy surfaces of the ${}^2\Delta(d^9s^1)$ (dashed line), ${}^4\Delta(d^8s^2)$ (dotted line), and ${}^2\Delta(d^8s^2)$ (solid line) states are given in Figure 1a. It is immediately obvious that the UHF method does not predict the correct ground state since both of the ${}^2,4\Delta(d^8s^2)$ states lie over 1 eV lower in energy than the ${}^2\Delta(d^9s^1)$ state. A stability analysis of the resulting UHF wave functions also reveals that only the lower energy ${}^2,4\Delta(d^8s^2)$ solutions are stable. *The experimentally observed ground state is unstable using the UHF method* due to the lack of electron correlation, which would preferentially stabilize configurations with a larger number of d electrons (an NBO analysis at $R = 1.75$ Å yields d populations of 8.0 and 8.9 electrons for the ${}^2\Delta(d^8s^2)$ and ${}^2\Delta(d^9s^1)$ states, respectively).

The calculated spin properties of the two UHF solutions are also dramatically different. For example, at a bond distance of 1.75 Å, the ${}^2\Delta(d^9s^1)$ “ground” state has an S^2 value of 0.76 (close to the expected value of 0.75) but lies much higher in energy. The ${}^2\Delta(d^8s^2)$ “excited” state has an

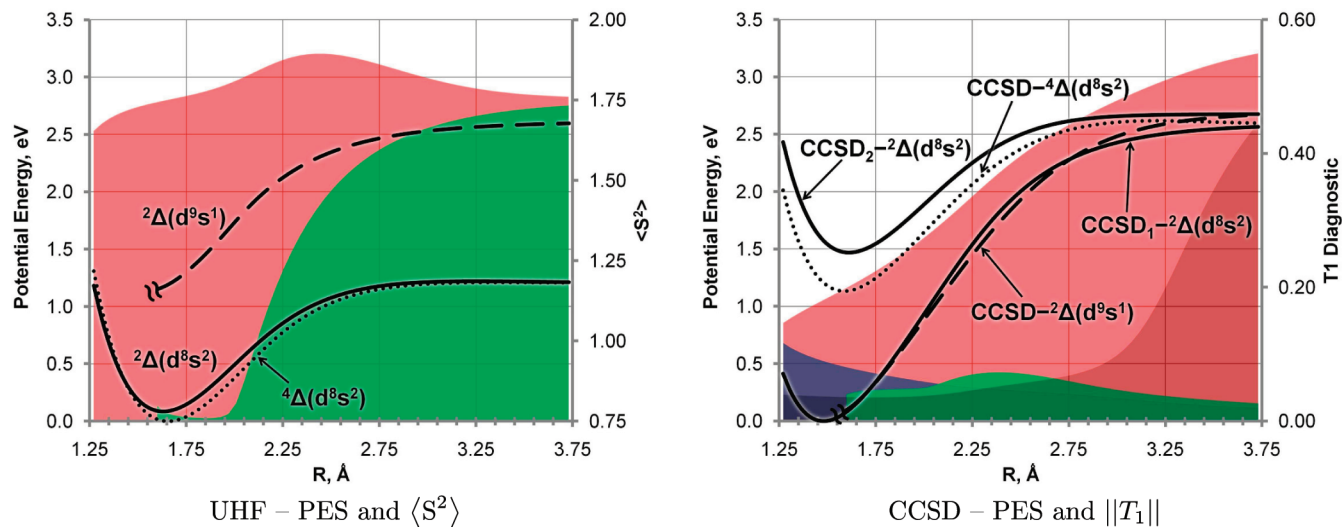


Figure 1. Potential energy surfaces. (a) UHF solutions for the “ground” ${}^2\Delta(d^9s^1)$ state (dashed line) and “excited” states ${}^2\Delta(d^8s^2)$ (solid line) and ${}^4\Delta(d^8s^2)$ (dotted line). PESs are shown as curves and plotted against the left axis. $\langle S^2 \rangle$ values are shown for doublet states as areas plotted against the right axis. Pink: $\langle S^2 \rangle$ for the ${}^2\Delta(d^8s^2)$ state. Green: $\langle S^2 \rangle$ for the ${}^2\Delta(d^9s^1)$ state. (b) Both unique CCSD₁ and CCSD₂ solutions for the ${}^2\Delta(d^8s^2)$ UHF reference state (solid line). The single CCSD solution for the ${}^2\Delta(d^9s^1)$ UHF reference state (dashed line). The single CCSD solution for the ${}^4\Delta(d^8s^2)$ UHF reference state (dotted line). $\|T_1\|$ values are shown as areas plotted against the right axis. Pink: $\|T_1\|$ for the CCSD₁ solution. Purple: $\|T_1\|$ for the CCSD₂ solution. Green: $\|T_1\|$ for the CCSD- ${}^2\Delta(d^9s^1)$ solution. Gray: $\|T_1\|$ for the CCSD- ${}^4\Delta(d^8s^2)$ solution.

S^2 value of 1.76 and lies lower in energy by 1.2 eV. As depicted in Scheme 1, this state has the following electron configuration $\delta_{3d}^{\alpha}\sigma_{3d}^{\alpha}\sigma_{4s}^{\beta}$. The corresponding $\langle S^2 \rangle$ values are shown as areas and plotted as a function of Ni–H bond length. The pink $\langle S^2 \rangle$ area shown in Figure 1a illustrates the extreme amount of spin contamination which is found to exist at all bond lengths for the stable ${}^2\Delta(d^8s^2)$ state, an observation which has been known for some time.¹⁹ The source of spin contamination is also clear from the similarity in the potential energy curves for the ${}^2,{}^4\Delta(d^8s^2)$ states. The result of this spin contamination is that the calculated bond length for the lowest energy ${}^2\Delta$ state at the UHF level (1.63 Å) is substantially larger than the ground state experimental value of 1.47 Å.³⁹ As expected, very little spin contamination is observed for the ${}^4\Delta(d^8s^2)$ UHF solution. Overall, the poor performance of UHF is striking.

2.2. Multiple CCSD Solutions. The different UHF solutions were then used to obtain potential energy surfaces (PES) at the CCSD level, leading to unanticipated results. Our most exciting result is that for a single UHF reference wave function, we found two different converged CCSD solutions. Starting from the stable UHF solution for the ${}^2\Delta(d^8s^2)$ state, we have been able to converge to two CCSD wave functions and energies. To the best of our knowledge, this is the first example of the existence of multiple solutions to the CCSD equations for a chemically interesting molecule with available experimental data. In Figure 1b, the two unique CCSD PESs using the same ${}^2\Delta(d^8s^2)$ UHF reference state are given as solid curves, while the CCSD solution found for the unstable ${}^2\Delta(d^9s^1)$ UHF reference state is given as a dashed line (note that since the UHF ${}^2\Delta(d^9s^1)$ state could not be found for small bond lengths, the CCSD curve also cannot be found). The quartet surface is represented with a dotted line. For all three PESs, we also plot, as areas, the T_1 diagnostic ($\|T_1\|$) of Lee and co-workers⁴⁰ as a function of the Ni–H bond

length. The pink, purple, green, and gray areas represent the $\|T_1\|$ values for the CCSD₁, CCSD₂, CCSD- ${}^2\Delta(d^9s^1)$, and CCSD- ${}^4\Delta(d^8s^2)$ wave functions, respectively.

As pointed out by previous authors studying the H4 model system,⁶ while finding one solution is typically easy, the others are often more difficult. The easily obtained solution is referred to as the “standard” solution, whereas the more difficult solutions are referred to as “alternate” solutions. Of the two CCSD solutions sharing the same UHF reference orbitals, CCSD₁ was found readily using the standard convergence algorithms (using coefficients from first-order perturbation theory (from an MP2 calculation) as the initial set of amplitudes). Therefore, CCSD₁ is considered our “standard” solution. To obtain CCSD₂, we first ran a CCD calculation to obtain a set of amplitudes that were expected to be closer to the converged CCSD amplitudes than the perturbation theory coefficients, though the orbital relaxation effects from the T_1 amplitudes are still neglected. Using the converged CCD amplitudes as our initial guess for the CCSD amplitudes, we were able to converge to a second CCSD solution (CCSD₂) at a stretched bond length. These converged amplitudes were then used as the initial amplitude guess for the next point on the potential energy surface. This was repeated to compute the full CCSD₂ PES. Both the “standard” solution and the “alternate” solution are shown in Figure 1b with solid black lines.

3. CCSD Solution—Electronic State Correspondence

Upon finding multiple CCSD solutions, one must address the following questions:

- Do the CC solutions correspond to actual physical states?
- Which solution corresponds to the same electronic state as the reference wave function?

Table 1. Calculated Spectroscopic Parameters for the ${}^2\Delta$ State

solution	reference	dissociation	R_{eq} (Å)	D_e (eV)
UHF ₁		$d^8s^2 + s^1$	1.630	1.12
CCSD ₂	${}^2\Delta(d^8s^2)$	$d^8s^2 + s^1$	1.611	1.21
CCSD ₁	${}^2\Delta(d^8s^2)$	$d^9s^1 + s^1$	1.477	2.71
UHF ₂		$d^9s^1 + s^1$		
CCSD	${}^2\Delta(d^9s^1)$	$d^9s^1 + s^1$		

• To which electronic states do the remaining solutions correspond?

In the absence of a mathematically rigorous way to obtain a physical state correspondence for the different CCSD solutions, we have investigated many aspects related to the CCSD wave functions to make our determinations. As the multiple CCSD solutions arise from a doublet UHF solution, we only concern ourselves with the doublet solutions for electronic state determination.

Considering that the PESs for both CCSD solutions appear reasonable on initial inspection (i.e., separation into atomic states, appropriate equilibrium bond lengths, and binding energies), it seems as though both solutions correspond to physical states. Simple analysis of the PESs shows that the “standard” solution (CCSD₁) actually corresponds to a different electronic state (${}^2\Delta(d^9s^1)$) than is described by the reference wave function (${}^2\Delta(d^8s^2)$), and the “alternate” solution (CCSD₂) corresponds to the same ${}^2\Delta(d^8s^2)$ state as the reference UHF wave function. Some of the characteristics of the different PESs are summarized in Table 1. As shown in this table, the UHF wave function values for R_{eq} and D_e are similar to CCSD₂, whereas the parameters for CCSD₁ are very different.

3.1. Spin Densities. Determining the electronic state for a single determinant wave function is normally rather straightforward via direct inspection of the molecular orbitals. However, this type of analysis is more complicated with a correlated, multideterminantal wave function. In Figure 2, we report isodensity plots of the spin densities ($\rho^\alpha - \rho^\beta$) using both the reference wave function density and the associated CCSD response density(ies). Blue surfaces indicate excess α density, and black surfaces indicate regions of excess β density. Therefore, a pure doublet spin state would have no visible black surface shown.

In Figure 2a and b, respectively, the UHF₂ and associated CCSD spin densities are shown. As shown in Figure 1a, the UHF₂ wave function is already rather close to a pure doublet ($\langle S^2 \rangle = 0.763$), so little is required of the T_1 amplitudes in terms of cleaning up the spin contamination, and thus the resulting CCSD spin density looks rather similar to the UHF spin density and the T_1 diagnostic is small ($\|T_1\| = 0.05$). Both the UHF₂ and resulting CCSD results clearly resemble ${}^2\Delta$ states, with a single unpaired electron occupying a d -type orbital.

In Figure 2c,d,e, spin densities are given for the UHF₁ reference wave function (d^8s^2) and associated CCSD₂ and CCSD₁ solutions. In addition to the similarities in the spectroscopic parameters shown in Table 1, inspection of the spin densities provides further evidence that the “alternate” solution (CCSD₂) more closely resembles the reference

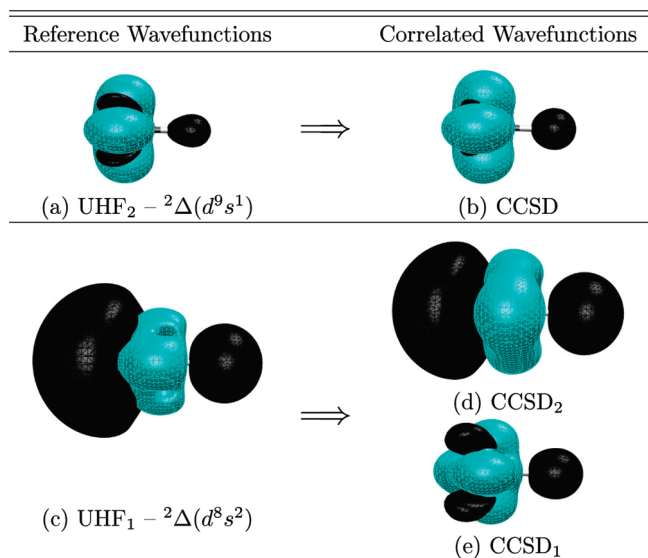


Figure 2. Spin density plots ($\rho^\alpha - \rho^\beta$, isovalue = 0.002) shown at a Ni–H bond length $R = 1.75$ Å. Excess α density (blue). Excess β density (black). Left column, a and c, shows the spin densities of the different reference wave function spin densities. Right column shows the different correlated wave function spin densities. (b) CCSD solution using the UHF₂ reference. (d and e) The two different CCSD solutions for the same reference (UHF₁).

wavefunction. The extreme amount of spin polarization resulting in the high degree of spin contamination can be immediately seen in the spin densities for UHF₁ and CCSD₂. It is most interesting that the “standard” solution (CCSD₁) starting from UHF₁ describes a different electronic state (d^9s^1) that more closely resembles UHF₂. These results are quite surprising as it means that the CCSD solution which was most easily obtained actually corresponds to a different electronic state than the original starting point. This reinforces the need to employ caution when studying highly correlated chemical systems, even with well-calibrated “black box” methods such as CCSD.

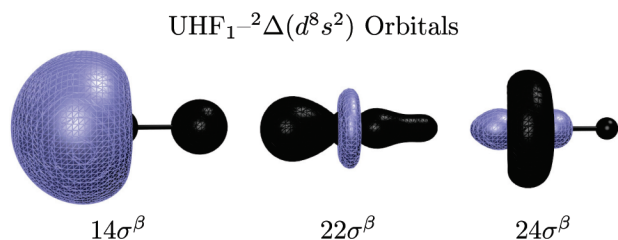
3.2. EOM-CCSD. To determine the electronic state correspondence for the “standard” solution (CCSD₁), we note the resemblance between the PES for CCSD₁ and the PES for the CCSD calculation starting from the UHF₂ reference (comparing regions of the PES for which both solutions could be obtained). This suggests that the CCSD₁ solution corresponds to the experimental ground state (and UHF₂ excited state) ${}^2\Delta(d^9s^1)$. This implies that the cluster operators within the spin-orbital formulation act upon the spin-symmetry-broken UHF₁ solution to produce a correlated wave function where the spin symmetry is mostly restored. However, the similarities in the energy alone are not sufficient to make our determination, as there are multiple low lying electronic states for NiH. To make a more definitive connection, we have calculated EOM-CCSD excited states using both the “standard” and “alternate” solutions, the results of which are listed in Table 2.

Table 2 shows the similarities in excitation energy between the lowest energy ${}^2\Delta$ excited states from both of the EOM-CCSD calculations. If we compute the excited states for the “standard” solution (CCSD₁), we find an excited ${}^2\Delta$ state

Table 2. EOM-CCSD Calculations with Both the Standard and Alternate CCSD Solutions^a

solution	lowest ${}^2\delta$ state (ev)	occ \rightarrow virt	coefficient
EOM-CCSD ₂	-1.327	$14\sigma^\beta\rightarrow 22\sigma^\beta$	-0.561
EOM-CCSD ₁	+1.343	$14\sigma^\beta\rightarrow 24\sigma^\beta$	+0.530
		$14\sigma^\beta\rightarrow 22\sigma^\beta$	+0.642
		$14\sigma^\beta\rightarrow 24\sigma^\beta$	-0.601

^a Both CCSD solutions use the UHF₁ (${}^2\Delta(d^8s^2)$) reference. Ni-H bond length, $R = 1.75$ Å.

**Figure 3.** Molecular orbitals which have significant contributions to the EOM-CCSD excited states.

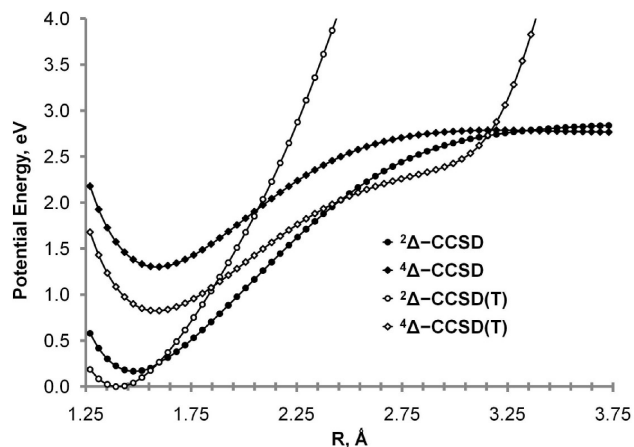
1.34 eV higher in energy. Likewise, if we calculate the excited states for the “alternate” solution (CCSD₂), we find that a ${}^2\Delta$ state lies 1.33 eV lower in energy. If we look at the results more closely, we see that for both excited states the largest coefficients are in front of determinants which involve switching the same UHF₁ orbitals (i.e., $14\sigma^\beta\rightarrow 22\sigma^\beta$ and $14\sigma^\beta\rightarrow 24\sigma^\beta$). These orbitals are shown in Figure 3.

As illustrated in Scheme 1, the ${}^2\Delta(d^9s^1)$ and ${}^2\Delta(d^8s^2)$ UHF solutions differ primarily in the β occupancy of the σ_{3d_z} and σ_{4s} orbitals. These are precisely the orbitals upon which a change in occupation connects the two excited ${}^2\Delta$ states in Table 2. It can be seen from Figure 3 that the molecular orbital $14\sigma^\beta$ is primarily of 4s character, and the $22\sigma^\beta$ and $24\sigma^\beta$ each have significant $3d_z$ character. We therefore conclude the following:

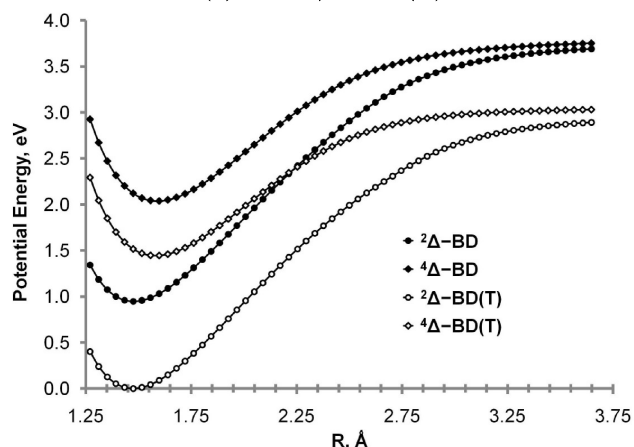
- Both the “standard” (CCSD₁) and “alternate” (CCSD₂) solutions correspond to physical states.
- The “standard” solution is assigned to the ${}^2\Delta(d^9s^1)$ state, and the “alternate” solution is assigned to the ${}^2\Delta(d^8s^2)$ state.
- The “alternate” solution describes the same electronic state as the reference UHF wave function.

4. Corrections to Correlated Wave Functions

As we have just noted, if one treats this system in a “black box” fashion and first obtains a stable ${}^2\Delta$ UHF solution, then employs standard convergence procedures to obtain a CCSD energy, the solution will have extremely large T_1 amplitudes related to the orbital rotations required to describe the ${}^2\Delta(d^9s^1)$ state from a ${}^2\Delta(d^8s^2)$ reference wave function. If one is only concerned with the CCSD energies, then this provides at least a qualitatively acceptable PES. However, often it is recognized that triple excitations are necessary to describe the system to a satisfactory degree of quantitative accuracy. This is commonly done via a perturbational correction to the correlated wave function using the well-known CCSD(T) method.^{41,42} We now explore some of the implications related to using CCSD(T) to describe an



(a) CCSD/CCSD(T)



(b) BD/BD(T)

Figure 4. Potential energy surfaces. (a) CCSD (solid line), CCSD(T) (dashed line). As this uses the ${}^2\Delta(d^8s^2)$ UHF reference wave function, the CCSD curve is the same as the CCSD₁ curve given in Figure 1. (b) BD (solid line), BD(T) (dashed line).

electronic state whose character is substantially different from the orbital occupations that define the reference wave function.

In Figure 4a, we show the lowest energy ${}^2\Delta$ and ${}^4\Delta$ potential energy curves for both CCSD and CCSD(T). While the CCSD curves are qualitatively correct, inclusion of triple excitations by means of a perturbational correction results in unphysical potential energy curves. This unrealistic rising of energy with the bond length is a direct result of the incredibly large T_1 amplitudes in the CCSD wave functions which are used to compute the triples correction. In Figure 1b, the T_1 diagnostic is given for the CCSD- ${}^2\Delta$ curve (pink) and the CCSD- ${}^4\Delta$ curve (gray). For both of these CCSD solutions, the $\|T_1\|$ becomes very large at long bond lengths. However, as a matrix norm, the $\|T_1\|$ metric does not fully indicate the nature of the individual T_1 amplitudes. Table 3 lists the largest amplitudes (greater than 0.2) for the CCSD₁ and CCSD₂ solutions. While there are many significantly large T_1 amplitudes, two particular components of the CCSD₁ wave function have coefficients greater than 1. The singly excited determinants $|\Psi_{14\sigma^\beta}^{22\sigma^\beta}\rangle$ and $|\Psi_{14\sigma^\beta}^{24\sigma^\beta}\rangle$ have coefficients of 1.05 and -1.01 , respectively. These incredibly large T_1

Table 3. Dominant Configurations in the CCSD₁ and CCSD₂ Wavefunctions^a

orbitals		amplitude
	CCSD ₁	
14σ ^β →22σ ^β		1.05
14σ ^β →24σ ^β		-1.01
14σ ^β →21σ ^β		-0.55
14σ ^β →27σ ^β		0.53
14σ ^β →18σ ^β		-0.45
14σ ^β →40σ ^β		-0.33
14σ ^β →30σ ^β		0.30
14σ ^β →37σ ^β		-0.26
14σ ^β →17σ ^β		0.20
	CCSD ₂	
14σ ^β →22σ ^β		-0.29
14σ ^β →24σ ^β		0.26

^aAll amplitudes greater than 0.2 are shown. Bond length, 1.75 Å.

amplitudes, required to obtain the ${}^2\Delta(d^9s^1)$ state from a ${}^2\Delta(d^8s^2)$ reference wave function, render any correction based on perturbation theory inappropriate. Note that these two excited determinants are obtained by swapping the exact same orbitals that are active in the EOM-CCSD calculations, as shown in Table 2.

If, however, one opts to use a different MO basis, one in which all the T_1 amplitudes are zero by design, better behavior might be expected. The BD method (Brueckner Doubles) is defined as a CCSD approach in which all of the T_1 amplitudes have been driven to zero via orbital rotations.⁴³ In Figure 4b, the BD and BD(T) potential energy curves are given for both the lowest energy ${}^2\Delta$ and ${}^4\Delta$ potential energy curves. While the BD and CCSD methods yield very similar potential energy curves, the triples correction for BD behaves far better. Because the T_1 amplitudes in a BD wave function are zero, the singles–doubles terms in the (T) correction are also zero, and thus there are no terms which may provide a destabilizing effect on the energy.

Note that Figure 4a shows CCSD underestimating the 1.57 eV ${}^2\Delta$ – ${}^4\Delta$ energy separation by 0.47 eV.³² Performance worsens significantly when the triples correction is added to CCSD (calculated energy separation of only 0.82 eV). As seen in Figure 4b, the performance of BD is very similar to that of CCSD (BD underestimates this state separation by 0.48 eV). However, due to the reasons just outlined, the BD wave function does not suffer from the accumulation of large T_1 amplitudes, and thus BD(T) performs quite well, yielding a ${}^2\Delta$ – ${}^4\Delta$ energy separation of 1.44 eV. This is in good agreement with the results from the multireference calculations.³² The calculated BD(T) bond lengths for the ${}^2\Delta$ and ${}^4\Delta$ states (1.479 Å and 1.587 Å) are also in good agreement with the multireference results.

The energy difference between the ${}^2\Delta(d^9s^1)$ and ${}^2\Delta(d^8s^2)$ states (experimentally measured to be 2.01 eV)³⁹ is more difficult to compute. At the CCSD level, the calculated difference is too small, 1.5 eV, since the ${}^2\Delta(d^9s^1)$ ground state is described poorly. However, since the ${}^2\Delta(d^8s^2)$ and ${}^4\Delta(d^8s^2)$ states are somewhat similar, the calculated energy difference between them at the CCSD level (0.3 eV) is likely to be more reasonable. If this energy difference is added to the computed ${}^2\Delta(d^9s^1)$ – ${}^4\Delta(d^8s^2)$ energy difference (1.44

eV at the BD(T) level), we determine the energy difference between the two lowest ${}^2\Delta$ states as 1.74 eV, in reasonable agreement with the experimental value of 2.01 eV.

5. Conclusions

In this article, we report the calculation of multiple solutions to the CCSD equations and subsequent analysis of the results, which suggest interesting implications for coupled cluster-based applications. From this work we have drawn the following conclusions:

1. The ability to find alternate solutions to the CCSD equations largely depends on the underlying wave function. For systems whose ground state orbital occupations change after the inclusion of electron correlation, one needs to be wary of the resulting wave function, as it may not correspond to the same electronic state. Analysis of the T_1 diagnostic is useful in determining if the correlated wave function is describing a different electronic state. This has implications for deliberately using unstable UHF wave functions as references in CCSD calculations for the purpose of modeling excited states.
2. EOM-CCSD is useful as a test to determine if lower energy CCSD solutions exist.
3. We have found that using the converged CCD amplitudes as a set of initial guess amplitudes for a CCSD calculation improves the ability to converge to a solution which describes the same state as the reference.
4. The use of a Brueckner orbital reference wave function reduces the possibility of converging to CCSD solutions which describe different electronic states than the reference.
5. In addition to other beneficial aspects, obtaining a correlated wave function from a set of Brueckner orbitals yields a correlated wave function which is much more appropriate for use in perturbative treatments for higher order corrections such as BD(T).
6. NiH is an interesting molecule which exhibits many characteristics which are often very difficult to describe theoretically. Thus, its use as a test molecule is further encouraged.

Acknowledgment. This work was supported by an NSF grant, CHE-0616737, at Indiana University.

References

- (1) Coester, F.; Kümmel, H. *Nucl. Phys.* **1960**, *17*, 477–485.
- (2) Cizek, J. *J. Chem. Phys.* **1966**, *45*, 4256.
- (3) Paldus, J.; Shavitt, I.; Cizek, J. *Phys. Rev. A* **1972**, *5*, 50.
- (4) Zivkovic, T. P.; Monkhorst, H. J. *J. Math. Phys.* **1978**, *19*, 1007.
- (5) Meissner, L.; Balkova, A.; Bartlett, R. J. *Chem. Phys. Lett.* **1993**, *212*, 177–184.
- (6) Jankowski, K.; Kowalski, K.; Jankowski, P. *Int. J. Quantum Chem.* **1994**, *50*, 353–367.
- (7) Jankowski, K.; Kowalski, K. *J. Chem. Phys.* **1999**, *110*, 3714–3729.

- (8) Jankowski, K.; Kowalski, K. *J. Chem. Phys.* **1999**, *110*, 9345–9352.
- (9) Jankowski, K.; Paldus, J. *Int. J. Quantum Chem.* **1980**, *18*, 1243.
- (10) Pariser, R.; Parr, R. G. *J. Chem. Phys.* **1953**, *21*, 466.
- (11) Pople, J. A. *Trans. Faraday Soc.* **1953**, *49*, 1375.
- (12) Lawler, K. V.; Parkhill, J. A.; Head-Gordon, M. *J. Chem. Phys.* **2009**, *130*, 184113.
- (13) Podeszwa, R. *Chem. Phys. Lett.* **2002**, *365*, 211–215.
- (14) Podeszwa, R.; Stolarczyk, L. *Chem. Phys. Lett.* **2002**, *366*, 426–432.
- (15) Podeszwa, R.; Stolarczyk, L. Z.; Jankowski, K.; Rubiniec, K. *Theor. Chem. Acc.* **2003**, *109*, 309.
- (16) Jankowski, K.; Kowalski, K. *J. Chem. Phys.* **1999**, *111*, 2940–2951.
- (17) Jankowski, K.; Kowalski, K.; Jankowski, P. *Int. J. Quantum Chem.* **1995**, *53*, 501–514.
- (18) Jankowski, K.; Kowalski, K.; Jankowski, P. *Chem. Phys. Lett.* **1994**, *222*, 608–614.
- (19) Guse, M. P.; Blint, R. J.; Kunz, A. B. *Int. J. Quantum Chem.* **1977**, *11*, 725–732.
- (20) Bagus, P. S.; Bjorkman, C. *Phys. Rev. A* **1981**, *23*, 461.
- (21) Walch, S. P.; Bauschlicher, C. W., Jr. *Chem. Phys. Lett.* **1982**, *86*, 66–70.
- (22) Blomberg, M. R. A.; Siegbahn, P. E. M.; Roos, B. O. *Mol. Phys.* **1982**, *47*, 127–143.
- (23) Walch, S. P.; Bauschlicher, C. W., Jr. *J. Chem. Phys.* **1983**, *78*, 4597.
- (24) Rohlfing, C. M.; Hay, P. J.; Martin, R. L. *J. Chem. Phys.* **1986**, *85*, 1447–1455.
- (25) Langhoff, S. R.; Bauschlicher, C. W., Jr. *Annu. Rev. Phys. Chem.* **1988**, *39*, 181–212.
- (26) Marian, C. M.; Blomberg, M. R. A.; Siegbahn, P. E. M. *J. Chem. Phys.* **1989**, *91*, 3589–3595.
- (27) Nelis, T.; Beaton, S. P.; Evenson, K. M.; Brown, J. M. *J. Mol. Spectrosc.* **1991**, *148*, 462–478.
- (28) Kadavathu, S. A.; Scullman, R.; Field, R. W.; Gray, J. A.; Li, M. *J. Mol. Spectrosc.* **1991**, *147*, 448–470.
- (29) Gray, J. A.; Li, M.; Nelis, T.; Field, R. W. *J. Chem. Phys.* **1991**, *95*, 7164.
- (30) Pou-Amerigo, R.; Merchan, M.; Nebot-Gil, I.; Malmqvist, P.; Roos, B. O. *J. Chem. Phys.* **1994**, *101*, 4893–4902.
- (31) Harrison, J. F. *Chem. Rev.* **2000**, *100*, 679–716.
- (32) Zou, W.; Liu, W. *J. Comput. Chem.* **2007**, *28*, 2286–2298.
- (33) Mayhall, N. J.; Raghavachari, K.; Redfern, P. C.; Curtiss, L. A.; Rassolov, V. *J. Chem. Phys.* **2008**, *128*, 144122.
- (34) Goel, S.; Masunov, A. E. *J. Chem. Phys.* **2008**, *129*, 214302.
- (35) Sonnenberg, J. L.; Schlegel, H. B.; Hratchian, H. P. In *Computational Inorganic and Bioinorganic Chemistry*; Solomon, E. I., Scott, R. A., King, R. B., Eds.; John Wiley & Sons, Ltd.: New York, 2009; Chapter Spin Contamination in Inorganic Chemistry Calculations, pp 173–186.
- (36) Raghavachari, K.; Trucks, G. W. *J. Chem. Phys.* **1989**, *91*, 1062.
- (37) Moore, C. E. Atomic Energy Levels; National Bureau of Standards (U.S.): Gaithersburg, MD, 1958; Circular No. 467, III.
- (38) As our focus is to address issues related to solving the CC equations in general and not necessarily matching experimentally obtained values, we have not included any relativistic corrections in our calculations. For correlated calculations, we correlate only the 3s3p3d4s electrons.
- (39) Huber, K.; Herzberg, G. *Molecular Spectra and Molecular Structure IV. Constants of Diatomic Molecules*; Van Nostrand: New York, 1979; pp 464–465.
- (40) Lee, T. J.; Rendell, A. P.; Taylor, P. R. *Int. J. Quantum Chem., Quant. Chem. Symp.* **1989**, *S23*, 199–207.
- (41) Raghavachari, K.; Trucks, G. W.; Pople, J. A.; Head-Gordon, M. *Chem. Phys. Lett.* **1989**, *157*, 479–483.
- (42) Stanton, J. F. *Chem. Phys. Lett.* **1997**, *281*, 130–134.
- (43) Handy, N. C.; Pople, J. A.; Head-Gordon, M.; Raghavachari, K.; Trucks, G. W. *Chem. Phys. Lett.* **1989**, *164*, 185–192.

CT100321K

Benchmark Calculations of Absolute Reduction Potential of Ferricenium/Ferrocene Couple in Nonaqueous Solutions

Mansoor Namazian,* Ching Yeh Lin, and Michelle L. Coote*

ARC Centre of Excellence for Free-Radical Chemistry and Biotechnology, Research School of Chemistry, Australian National University, Canberra ACT 0200, Australia

Received June 15, 2010

Abstract: High-level ab initio molecular orbital theory is used to obtain benchmark values for the ferricenium/ferrocene (Fc^+/Fc) couple, the IUPAC recommended reference electrode for nonaqueous solution. The gas-phase ionization energy of ferrocene is calculated using the high-level composite method, G3(MP2)-RAD, and two higher-level variants of this method. These latter methods incorporate corrections for core correlation and, in the case of the highest level considered, use (RO)CCSD(T)/6-311+G(d,p) in place of (RO)CCSD(T)/6-31G(d) as the base level of theory. All methods provide good agreement with one another and the corresponding experimental values. Solvation energies have been calculated using PCM, CPCM, SMD, and COSMO-RS. Using G3(MP2)-RAD-Full-TZ gas-phase energies and COSMO-RS solvation energies, the absolute redox potentials of the Fc^+/Fc couple have been calculated as 4.988, 4.927, and 5.043 V in acetonitrile, 1,2-dichloroethane, and dimethylsulfoxide solutions, respectively.

1. Introduction

In contrast to aqueous solution, the method of measuring electrode potentials has not been well established in nonaqueous solutions.¹ One of the serious problems faced is the choice of the reference electrode. For nonaqueous solutions, there is no primary reference electrode equivalent to the aqueous standard hydrogen electrode (SHE) and no general reference electrode as reliable as the aqueous reference electrodes. Although aqueous reference electrodes are often used for nonaqueous systems, the liquid junction potential (LJP) between the aqueous and nonaqueous solutions can affect the measured potentials.² As a result, the IUPAC Commission on Electrochemistry has proposed that the ferricenium/ferrocene (Fc^+/Fc) couple be used an internal reference for reporting electrode potentials in nonaqueous solutions.³

Recently, theoretical investigation of redox potentials of compounds in aqueous and nonaqueous solutions has attracted attention.^{4–12} Computational chemistry offers an attractive alternative to experimentation, particularly in

situations when experimental measurements are difficult due to the participation of other chemical reactions, or when it is necessary to clarify the role of individual reactions involved in the electrochemical processes. To date, the calculated redox potentials of nonaqueous solutions are typically reported versus an aqueous reference electrode and compared with the available experimental values.^{5,6} To be consistent with IUPAC recommendations, however, the calculated electrode potentials should be reported versus the Fc^+/Fc couple. Therefore, knowledge of the absolute reduction potential of this couple in nonaqueous solutions is necessary. Su and Girault have reported a value of 5.01 V for the absolute reduction potential for the Fc^+/Fc couple in 1,2-dichloroethane,¹⁰ which was compared with the value of 5.08 V as obtained from the sum of the corresponding aqueous SHE potential (4.44 V) and a correction of 0.64 V. It is worth noting that the absolute reduction potential of SHE might be different than 4.44 V, as it has been recently estimated as 4.24 and 4.27 V.^{7,13} In any case, the use of aqueous values of the SHE and SCE potentials to reproduce the redox potential of the Fc^+/Fc couple in a nonaqueous solution is problematic. Computationally, the *relative* redox potential of ferrocene in acetonitrile has been studied by Baik

* Corresponding authors. E-mail: namazian@rsc.anu.edu.au (M.N.), mcoote@rsc.anu.edu.au (M.L.C.).

Table 1. Principal Geometric Parameters of Fc

bond	bond length (Å)					
	LanL2DZ ^a	TZQ ^a	aug-cc-pVTZ-DK ^a	LanL2TZf ^a	LanL2TZf ^b	Experiment ³⁴
Fe–Cp	1.69	1.68	1.68	1.68	1.68	1.66
Fe–C	2.08	2.07	2.08	2.07	2.07	2.06
C–C	1.43	1.43	1.43	1.43	1.43	1.44
C–H	1.08	1.08	1.08	1.08	1.08	1.10

^a 6-31G(d) basis set has been used for all H and C atoms. ^b 6-311+G(d,p) basis set has been used for all H and C atoms.

and Friesner,¹⁴ who used the Saturated Calomel Electrode (SCE) as the reference electrode, and Roy et al.,¹⁵ who used SHE as the reference electrode. However, converting these values to the corresponding *absolute* potentials is again hampered by the difficulty in estimating the relevant liquid junction potential. Moreover, the absolute values of the reduction potentials for the Fc⁺/Fc couple in most other nonaqueous solutions are not known, and relating theoretical and experimental redox potentials is an ongoing problem.

In the present work, we use high-level *ab initio* molecular orbital theory to calculate an accurate value for the absolute redox potential of the Fc⁺/Fc couple in several common nonaqueous solvents: acetonitrile (AN), 1,2-dichloroethane (DCE), and dimethylsulfoxide (DMSO). The accurate theoretical values reported here can be used widely in order to calculate the relative reduction potential of other species vs the Fc⁺/Fc couple in nonaqueous solution.

2. Computational Methods

The geometries of studied species were optimized at the B3-LYP level of theory using the LanL2DZ,¹⁶ LanL2TZf,^{17,18} TZQ,¹⁹ and aug-cc-pVTZ-DK²⁰ basis sets for the Fe atom and the 6-31G(d) and 6-311+G(d,p) basis sets for C and H. Single-point energies were calculated using the high-level composite method G3(MP2)-RAD.^{21,22} This method approximates (RO)CCSD(T, FC) with the large triple- ζ basis set GTMP2large as the sum of the corresponding (RO)CCSD(T, FC)/6-31G(d) calculations and a basis set correction term, evaluated at the ROMP2 level of theory. This procedure is normally considered to achieve “chemical accuracy” (ca. 0.05 eV) for gas-phase organic thermochemistry.^{21,22} However, since the present compounds contain a transition metal, we also considered two improved versions of this method. In the first, which we refer to as G3(MP2)-RAD-Full, we added an additional correction for core correlation, evaluated as the difference of corresponding calculations at the (RO)CCSD(T, Full)/6-31G(d) and (RO)CCSD(T, FC)/6-31G(d) levels. In the second, we used (RO)CCSD(T, FC)/6-311+G(d,p) as our highest level of theory, so that the ROMP2 basis set correction to GTMP2large was much less significant. We refer to this method, which also included the core correlation corrections, as G3(MP2)-RAD-Full-TZ. Further details, including all component calculations, are provided in the Supporting Information.

Gas-phase zero-point energies, thermal corrections, and entropic corrections were calculated using the standard formulas for the statistical thermodynamics of an ideal gas under the harmonic oscillator approximation using the optimized geometries and scaled²³ B3-LYP/6-31G(d)/

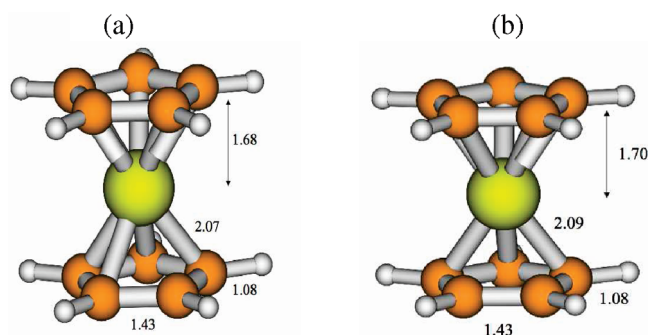


Figure 1. Optimized structure of (a) ferrocene and (b) the ferricinium ion.

LanL2TZf frequencies. However, for the low barrier rotation of the Cp rings, the thermal and entropic corrections were calculated using the standard free rotor model formulas. Solvation energies of the studied species in the various solvents were calculated using PCM and CPCM continuum models^{24,25} calculated using UAKS radii at the recommended²⁶ level of theory, B3LYP/6-31+G(d,p). Calculations were also performed using the recently introduced solvation models, SMD²⁷ and COSMO-RS,^{28,29} at the B3LYP/6-31G(d) and BP/TZP levels of theory, respectively. In all solvation energy calculations, the LanL2TZf basis set was used for Fe. The default values of Klamt^{28,29} and the SMD-Coulomb atomic radii²⁷ have been used for COSMO-RS and SMD, respectively. Since Fe was not present in the original parametrization sets for either of these models, the default SMD-Coulomb settings revert to the UAKS radius for Fe (1.456 Å), while the default COSMO-RS settings revert to the Allinger radius for Fe (1.858 Å). While neither of these values have been optimized for their respective models, this is not likely to cause a significant error in this system since the Fe atom is shielded by the two ligands.

All calculations were carried out using Gaussian 03³⁰ except for the (RO)CCSD(T) calculations, which were carried out using Gaussian 09,³¹ and the COSMO-RS calculations, which were performed using ADF^{32,33} software.

3. Results and Discussion

Geometries. Table 1 shows the optimized selected bond lengths of Fc using different basis sets for the Fe atom along with the corresponding literature values.³⁴ The geometry is relatively insensitive to the level of theory used, though, not surprisingly, the larger basis sets for Fe give slightly better results for the Cp–Fe distance than LanL2DZ. Figure 1 shows the optimized geometry of Fc and Fc⁺ calculated at the level of B3-LYP using LanL2TZf for Fe and 6-31G(d) for H and C atoms. As shown in this figure, the bond length

Table 2. Adiabatic Ionization Energy of Fc^a

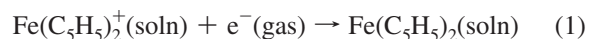
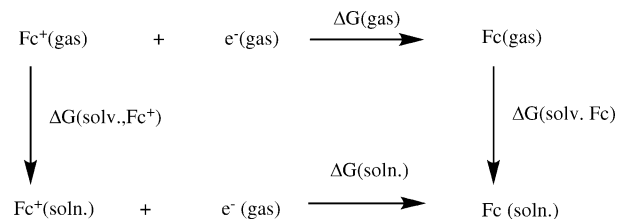
method	IE (0K) eV
G3(MP2)-RAD	7.062
G3(MP2)-RAD-Full ^b	7.067
G3(MP2)-RAD-Full-TZ ^{b,c}	7.047
exp. values ³⁶	6.6–7.2

^a Single-point energy calculations performed using the geometries optimized at the B3-LYP/6-31G(d)/LanL2TZf level. Unless otherwise noted, the triple- ζ LanL2TZf basis set is used for Fe in all of the improved energy calculations. Computational results are shown to the nearest 0.001 eV to indicate the level of precision in the calculations; the likely level of accuracy is ca. 0.05 eV. ^b Includes an additional correction for core correlation, using the (RO)CCSD(T,Full)/6-31G(d)/LanL2TZf level of theory. ^c (RO)CCSD(T)/6-311+G(d,p) has been used as the base level of theory in this method; see the Computational Methods section and Supporting Information for more details.

of Fe–C in Fc⁺, 2.09 Å, is longer than the corresponding bond length in Fc, 2.07 Å, which is in perfect agreement with the experiment.³⁴ Thus, removal of an electron from the bonding orbital of Fc leads to an increase of the Fe–Cp distance, suggesting that the strength of the Fe–C bond has been reduced. Although one might expect that the staggered isomer (*D*_{5d}) is more favorable, the equilibrium conformation in the gas phase is the eclipsed (*D*_{5h}) form, and this is in agreement with previous experimental and theoretical studies.³⁵

Adiabatic Ionization Energy. To calculate the redox potential for the Fc⁺/Fc couple, we first require the gas-phase ionization energy of Fc. Surprisingly, the literature values reported for the ionization energy of Fc cover a wide range, from 6.6 to 7.2 eV (a difference of 0.6 eV or 16 kcal mol⁻¹).³⁶ We first used the high-level composite ab initio method G3(MP2)-RAD to calculate the adiabatic ionization energy (see Table 2). In applying this method, we used the triple- ζ basis set LanL2TZf for Fe in all steps of the calculation. As noted in the Computational Methods section, we also considered two improved versions of this method, G3(MP2)-RAD-Full and G3(MP2)-RAD-Full-TZ. Both of these methods include an additional correction for core correlation, and the latter method also uses (RO)CCSD(T, FC)/6-311+G(d,p) as the base level of theory in place of (RO)CCSD(T, FC)/6-31G(d). These improved procedures deliver results that differ from the standard G3(MP2)-RAD calculations by 0.015 eV or less, which provides a good indication that the results have converged. In particular, it is worth noting that the effect of core correlation, often important for a complete description of transition metal chemistry, is negligible (ca. 0.005 eV) for this system. The G3(MP2)-RAD-Full-TZ ionization energy (IE = 7.046 eV), which is used for the remainder of this work, falls well within the scatter of the experimental values (6.6–7.2 eV) but is slightly higher than the currently recommended values of 6.71 ± 0.08 eV and 6.81 ± 0.07 eV.^{36,37}

Standard Reduction Potential of Fc⁺/Fc Couple. Upon reduction, the ferricinium radical cation gains one electron and converts into its reduced form, ferrocene:

**Scheme 1.** Thermodynamic Cycle Used to Calculate Gibbs Free Energy of Reaction 1

The total change in the Gibbs free energy of reaction 1 in solution, $\Delta G^\circ(\text{soln})$, is related to E° according to eq 2:³⁸

$$E^\circ = -\Delta G^\circ(\text{soln})/nF \quad (2)$$

where n is number of electrons transferred ($n = 1$ in this case) and F is the Faraday constant (23.061 kcal mol⁻¹ V⁻¹ or 96 485 C mol⁻¹).³⁸ To calculate $\Delta G^\circ(\text{soln})$, we have used the thermodynamic cycle shown in Scheme 1, which results in eq 3.

$$\Delta G^\circ(\text{soln}) = \Delta G^\circ(\text{gas}) + \Delta G^\circ(\text{solv. Fc}) - \Delta G^\circ(\text{solv. Fc}^+) \quad (3)$$

where $\Delta G^\circ(\text{gas})$ is the change of standard Gibbs free of reaction 1 in the gas phase, and $\Delta G^\circ(\text{solv. Fc})$ and $\Delta G^\circ(\text{solv. Fc}^+)$ are standard solvation energies of Fc and Fc⁺, respectively. $\Delta G^\circ(\text{gas})$ can be calculated using the adiabatic IE of Fc via eq 4:

$$\Delta G^\circ(\text{gas}) = \Delta H^\circ(\text{gas}) - T\Delta S(\text{gas}) = -\text{IE} + \text{TC} - T\Delta S(\text{gas}) \quad (4)$$

where TC is the thermal correction to the enthalpy and $\Delta S(\text{gas})$ is the change of entropy of the studied reaction. The thermal corrections and entropies have been calculated at the B3LYP/6-31G(d)/LanL2TZf level, and the results are tabulated in the Supporting Information. The contribution of $T\Delta S(\text{gas})$ is -0.108 eV, including the entropy of free electrons (5.43 cal mol⁻¹ K⁻¹),³⁹ and the contribution of TC has been calculated as -0.040 eV, including the correction for the enthalpy of free electrons³⁹ at 298 K; therefore, $\Delta G^\circ(\text{gas})$ is -6.979 eV or -160.9 kcal mol⁻¹. If we instead exclude the enthalpy and entropy of free electrons, following the “ion convention” (IC),³⁹ a value of -7.016 eV or -161.8 kcal mol⁻¹ for $\Delta G^\circ(\text{gas})$ of the reaction is obtained instead. The difference between these two results is negligible; for the remainder of this work, we have adopted the electron convention based on Fermi–Dirac statistics (EC-FD), recommended by Bartmess.³⁹

As shown by eq 3, Gibbs energies of solvation of both Fc and Fc⁺ are required in order to calculate the total change of Gibbs energy of reaction 1. Solvation energies are calculated using PCM and CPCM models of solvation together with recent models of COSMO-RS and SMD.^{26–29} The results, which are summarized in Table 3, show that PCM and CPCM models predict solvation energies for Fc in acetonitrile that are positive. Since Fc dissolves in acetonitrile, this is not physically realistic. In contrast, the calculated solvation energies of Fc obtained by COSMO-RS (-7.47 kcal mol⁻¹) and SMD (-8.50 kcal mol⁻¹) are

Table 3. Solvation Energies of Fc and Fc⁺ in Nonaqueous Solutions of Acetonitrile (AN), 1,2-Dichloroethane (DCE), and Dimethylsulfoxide (DMSO)

solvent	model	$\Delta G^\circ(\text{solv.})^a$ kcal mol ⁻¹	
		Fc	Fc ⁺
AN	PCM	+1.31	-40.37
AN	CPCM	+1.25	-40.40
AN	COSMO-RS	-7.47	-53.38
AN	SMD	-8.50	-54.97
AN	experiment ⁴⁰	-7.65	
DCE	SMD	-8.73	-52.01
DMSO	SMD	-6.03	-52.77

^a Solvation energies for PCM and CPCM have been calculated using UAKS radii at the B3LYP/6-31+G(d,p)/LanL2TZf level and for SMD and COSMO-RS using their respective default radii at the B3LYP/6-31G(d)/LanL2TZf and BP/TZP levels, respectively.

Table 4. Absolute Reduction Potentials (V) of the Fc⁺/Fc Couple in Nonaqueous Solution^a

solvent model	solvent		
	AN	DCE	DMSO
COSMO-RS	4.988	4.927	5.043
SMD	4.964	5.102	4.952
experiment	4.980 ^b	5.01 ^c	

^a Calculated at the G3(MP2)-RAD-Full-TZ level using various solvation methods as shown. Computational results are shown to the nearest 0.001 V to indicate the level of precision in the calculations; the likely level of accuracy is ca. 0.05–0.1 V. ^b Calculated from the experimental⁴¹ value of 0.380 V for the reduction potential of Fc⁺/Fc in acetonitrile relative to the Saturated Calomel Electrode (SCE), taking into account a recent value⁹ of 4.60 V for SCE in acetonitrile. ^c From ref 10.

negative and are in excellent agreement with experimental results (-7.65 kcal mol⁻¹).⁴⁰

Using G3(MP2)-RAD gas-phase energies together with the COSMO-RS and SMD solvation energies, the absolute values of the reduction potential of the Fc⁺/Fc couple in AN, DCE, and DMSO have been calculated and are shown in Table 4. Both sets of values are in excellent agreement with the available experimental data,^{10,41} with the COSMO-RS results marginally closer overall. As a final test of these numbers, we have previously used G3(MP2)-RAD to calculate the absolute redox potential of *para*-benzoquinone in acetonitrile.⁹ This absolute value (4.04 V) can now be combined with our new absolute value of the Fc⁺/Fc couple in the same solvent to obtain a theoretical value for the relative redox potential of this compound in acetonitrile. The values obtained (-0.948 or -0.924 V, depending on whether the SMD or COSMO-RS results are used for the reference electrode) both compare well with the experimental value for the same system (-0.851 V),⁸ with relatively small deviations of 0.097 and 0.073 V, respectively.

4. Conclusion

The absolute redox potentials of the Fc⁺/Fc couple in nonaqueous solutions of AN, DCE, and DMSO have been calculated as 4.988, 4.927, and 5.043 V and benchmarked against available experimental data for all components of the calculation. These values will allow for the calculation of the redox potentials of other species relative to the Fc⁺/Fc reference couple in most common nonaqueous solutions.⁴²

Acknowledgment. We gratefully acknowledge generous allocations of computing from the Australian National Computational Infrastructure and funding from the Australian Research Council under their Centres of Excellence program.

Supporting Information Available: B3-LYP/6-31G(d)/LanL2TZf optimized geometries and corresponding total energies, thermal corrections, and entropies at the various levels of theory studied. This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- (1) Kosuke, I. *Electrochemistry in Nonaqueous Solutions*; Wiley-VCH: New York, 2002; p 169.
- (2) Diggle, J. W.; Parker, A. J. *Aust. J. Chem.* **1974**, *27*, 1617.
- (3) Gritzner, G.; Kuta, J. *Pure Appl. Chem.* **1984**, *4*, 462.
- (4) Fu, Y.; Liu, L.; Yu, H.-Z.; Wang, Y.-M.; Guo, Q.-X. *J. Am. Chem. Soc.* **2005**, *127*, 7227.
- (5) Fu, Y.; Liu, L.; Wang, Y.-M.; Li, J.-N.; Yu, T.-Q.; Guo, Q.-X. *J. Phys. Chem. A* **2006**, *110*, 5874.
- (6) Namazian, M.; Norouzi, P. *J. Electroanal. Chem.* **2004**, *573*, 49.
- (7) Kelly, C. K.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. B* **2007**, *111*, 408.
- (8) Frontana, C.; Vázquez-Mayagoitia, A.; Garza, J.; Vargas, R.; González, I. *J. Phys. Chem. A* **2006**, *110*, 9411.
- (9) Namazian, M.; Coote, M. L. *J. Phys. Chem. A* **2007**, *111*, 7227.
- (10) Su, B.; Girault, H. H. *J. Phys. Chem B* **2005**, *109*, 11427.
- (11) Winget, P.; Cramer, C. J.; Truhlar, D. G. *Theor. Chem. Acc.* **2004**, *112*, 217.
- (12) Lewis, A.; Bumpus, A.; Truhlar, D. G.; Cramer, C. J. *J. Chem. Educ.* **2004**, *81*, 596.
- (13) Isse, A. A.; Gennaro, A. *J. Phys. Chem. B* **2010**, *114*, 7894.
- (14) Baik, M.-H.; Friesner, R. A. *J. Chem. Phys. A* **2002**, *106*, 4707.
- (15) Roy, L. E.; Jakubikova, E.; Graham Guthrie, M.; Batista, E. R. *J. Phys. Chem. A* **2009**, *113*, 6745.
- (16) Hay, P. J.; Wadt, W. R. *J. Chem. Phys.* **1985**, *82*, 270.
- (17) Hay, P. J.; Wadt, W. R. *J. Chem. Phys.* **1985**, *82*, 299.
- (18) Ehlers, W.; Bohme, M.; Dapprich, S.; Gobbi, A.; Hollwarth, A.; Jonas, V.; Kohler, K. F.; Stegmann, R.; Veldkamp, A.; Frenking, G. *Chem. Phys. Lett.* **1993**, *208*, 111.
- (19) Schultz, N. E.; Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2005**, *109*, 4388.
- (20) Balabanov, N.; Peterson, K. A. *J. Chem. Phys.* **2006**, *125*, 074110.
- (21) Henry, D. J.; Sullivan, M. B.; Radom, L. *J. Chem. Phys.* **2003**, *118*, 4849.
- (22) Henry, D. J.; Parkinson, C. J.; Radom, L. *J. Phys. Chem. A* **2002**, *106*, 7927.
- (23) Merrick, J. P.; Moran, D.; Radom, L. *J. Phys. Chem. A* **2007**, *111*, 11683.
- (24) Cossi, M.; Rega, N.; Scalmani, G.; Barone, V. *J. Comput. Chem.* **2003**, *24*, 669.
- (25) Barone, V.; Cossi, M. *J. Phys. Chem. A* **1998**, *102*, 1995.

- (26) Takano, Y.; Houk, K. N. *J. Chem. Theory Comput.* **2005**, *1*, 70.
- (27) Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. B* **2009**, *113*, 6378.
- (28) Klamt, A. *J. Phys. Chem.* **1995**, *99*, 2224.
- (29) Klamt, A.; Jonas, V.; Burger, T.; Lohrenz, J. C. W. *J. Phys. Chem. A* **1998**, *102*, 5074.
- (30) 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, 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.; Li, G.; Liu, A. L.; 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.
- (31) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, N. J.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, Ö.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. *Gaussian 09*; Gaussian, Inc.: Wallingford, CT, 2009.
- (32) Pye, C. C.; Ziegler, T.; van Lenthe, E.; Louwen, J. N. *Can. J. Chem.* **2009**, *87*, 790.
- (33) *ADF Program System*, release 2008.01; Scientific Computing & Modelling NV, Vrije Universiteit: Amsterdam, The Netherlands.
- (34) Haaland, A. *Acc. Chem. Res.* **1979**, *12*, 415.
- (35) Mayor-Lopez, M. J.; Weber, J. *Chem. Phys. Lett.* **1997**, *281*, 226.
- (36) *NIST Standard Reference Database Number 69*; Linstrom, P. J., Mallard, W. G., Eds.; National Institute of Standards and Technology: Gaithersburg, MD. <http://webbook.nist.gov> (retrieved March 1, 2010).
- (37) Meot-Ner, M. *J. Am. Chem. Soc.* **1989**, *111*, 2830.
- (38) Zare, H. R.; Eslami, M.; Namazian, M.; Coote, M. L. *J. Phys. Chem. B* **2009**, *113*, 8080.
- (39) Bartmess, J. E. *J. Phys. Chem.* **1994**, *98*, 6420.
- (40) Kuznetsov, A. M.; Maslii, A. N.; Krishtalik, L. I. *Russ. J. Electrochem.* **2008**, *44*, 34.
- (41) Pavlishchuk, V. V.; Addison, A. W. *Inorg. Chim. Acta* **2000**, *298*, 97.
- (42) Bordwell, F. G.; Harrelson Jr, J. A.; Satish, A. V. *J. Org. Chem.* **1989**, *54*, 3101.

CT1003252

Describing Anions by Density Functional Theory: Fractional Electron Affinity

Frank Jensen*

*Department of Chemistry, University of Aarhus, Langelandsgade 140,
DK-8000 Aarhus, Denmark*

Received June 16, 2010

Abstract: It is shown that the majority of commonly used exchange-correlation potentials in Kohn–Sham density functional theory describe anions as having only a fraction of the extra electron bound, while the remaining fraction drifts off to infinity when sufficiently flexible basis sets are employed. For systems with both cationic and anionic sites, this leads to fractional electron transfer, even when standard basis sets are used, and thus a qualitative incorrect description of the electronic structure. Exchange functionals without Hartree–Fock exchange display the largest effects, but hybrid functionals also show the phenomenon, except for strongly bound anions. The source of the error is the incorrect long-range behavior of the exchange potential and can be avoided by employing the long-range correction method. The results have consequences for density functional descriptions of systems with localized anionic or strong electron donor sites for almost all of the commonly employed exchange-correlation functionals.

I. Introduction

Density functional theory (DFT) in the Kohn–Sham version is firmly established as a useful tool for calculating a variety of properties for both molecular and extended systems.¹ The Hohenberg–Kohn theorem establishes that DFT is capable of providing an exact description of the electronic structure using only the electron density as a variable,² but at present only approximations to the elusive exchange-correlation (XC) potential exist. Different XC functionals are counted by the hundreds, and calibration of these functionals with various basis sets is a popular topic in the literature. For routine applications in molecular systems, the ubiquitous B3LYP functional³ is often the method of choice in combination with a medium sized Pople type basis set, like 6-31G*.⁴ In many cases such calculations provide useful results at a low computational cost, which has led to their widespread popularity.

One of the challenging properties to calculate accurately is the electron affinity (EA), corresponding to the energy difference between an atom or molecule and its corresponding anion. Sophisticated wave function methods are capable of achieving accuracies of a few milli-eV (few kJ/mol),⁵ but

such methods are computationally expensive. In the dawn of the DFT era in computational chemistry, concerns were raised that approximate DFT might not be suitable for describing anions, as the highest occupied molecular orbital (HOMO) energy often was calculated to be positive.^{6–8} A positive orbital energy describes an unbound electron, although the use of a finite basis set prevents this from happening in normal calculations. Systems with a formally unbound electron were nevertheless often found to have a positive EA when calculated as a difference between total energies of the anion and neutral species. Galbraith and Schaefer addressed these concerns by showing that the calculated EA for the fluorine atom does not change substantially by addition of multiple diffuse p-functions (smallest exponent used was $2 \cdot 10^{-5}$) with a selection of different XC functionals.^{9,10} The calculated EA with the B3LYP functional was 340 kJ/mol, and the HOMO energy for the fluorine anion converged to a value of essentially zero. The calculated EA with the BLYP functional was even larger, 355 kJ/mol, despite the fact that the HOMO energy for the anion was positive with a value of 0.053 au. Jarecki and Davidson latter reinvestigated the fluorine anion with the BLYP functional¹¹ and showed that by adding further diffuse functions (smallest exponent used was $7 \cdot 10^{-9}$) the EA increases further by ~ 15 kJ/mol and resulted in the

* Corresponding author e-mail: frj@chem.au.dk.

Table 1. Calculated Electron Affinities in kJ/mol with Four Different Methods^c

anion	HF		BHHLYP		B3LYP		BLYP		exp ^b
	apc-2	limit	apc-2	limit	apc-2	limit	apc-2	limit	
H ⁻	-27.4	0	64.3	68.8	88.2	105.1	82.2	116.4	72.8
Li ⁻	-10.8	0	41.9	44.2	53.6	62.3	43.7	63.1	59.6
Be ⁻	-88.3	0	-68.5	6.5	-51.8	27.8	-58.8	36.0	0
B ⁻	-26.2	0	18.8	34.9	46.9	74.7	45.6	89.4	27.0
C ⁻	53.6	53.3	104.1	104.1	136.7	150.7	135.2	168.2	121.9
N ⁻	-160.1	0	-28.4	21.4	21.1	77.0	35.2	108.2	0
O ⁻	-52.6	0	109.0	108.5	164.9	180.5	179.6	216.4	141.0
F ⁻	127.6	126.3	288.6	287.6	347.9	347.3	362.5	373.0	320.2
MAD^a	113.1	93.8	19.3	18.4	18.0	29.7	23.0	47.3	

^a Mean absolute deviation relative to experimental values for the six nonzero EAs. ^b Reference 13. ^c apc-2 indicates results obtained with the aug-pc-2 basis set; limit indicates the basis set limiting results.

HOMO energy stabilizing to a value of -0.065 au, indicating that the extra electron is bound when sufficiently diffuse functions are included in the basis set. We encountered the problem of formally unbound electrons when designing diffuse augmentation functions for the polarization consistent basis sets,¹² which are optimized for DFT calculations. Attempts of determining the optimum diffuse exponents by minimizing the BLYP energy for anions in most cases lead to the exponent diverging toward zero, and we suggested that only systems with a large EA may have well-defined EA values.

The early reservations regarding the ability of DFT to describe anions seem largely to have been overlooked or ignored by the practitioners in the field, and well over 1000 publications have appeared using DFT methods to calculate EA.^{13,14} In a few cases, the papers by Galbraith and Schaefer,⁹ and by Jarecki and Davidson,¹¹ are taken as evidence that DFT is capable of describing anions as stable species when sufficiently diffuse functions are included, although this is almost never done in actual calculations. The fluorine atom is one of the species with the largest known EA, and the above finding may thus not be generally applicable, but very few detailed investigations of other systems have been reported. It has been argued that atomic systems having strongly localized electrons will be the most problematic cases, and molecular anion where the extra electron can be delocalized over a larger volume of space will be better behaved,¹⁰ but again no detailed investigation has been reported. Even for the fluorine atom with a large EA, the puzzle remains how an anion with a positive HOMO energy and a formally unbound electron nevertheless can have a positive EA when calculated as a difference in total energies.

We will in the present paper show that most commonly employed XC functionals lead to DFT descriptions of anions as having only a fraction of the extra electron bound and that this is the rule rather than the exception for anions in general. We will use the term “fractional electron affinity” to describe this situation and show that this has consequences also for intermolecular ion complexes, zwitterionic and electron donor–acceptor systems in general.

II. Computational Details

Performing calculations with very diffuse basis functions faces a number of numerical issues that standard settings of

many commonly used program packages have not been designed for. All integral thresholds have been tightened to essentially machine precision, and density matrices have been converged to at least 10^{-8} . The XC contribution has been calculated using an Euler-Maclaurin radial grid¹⁵ with 5000 points in combination with a Lebedev angular grid with 434 points.¹⁶ A large radial grid is necessary for integrating electron density corresponding to basis functions with very small exponents, and the Davidson radial norm criterion¹¹ has in all cases been fulfilled to within 10^{-6} . It was checked that the results did not change upon further enlargement of the grid size. Threshold screenings for discarding integration points with low density for determining the XC energy have been disabled. Linear dependency of the basis sets is not a problem, as only a single expansion center is employed for the very diffuse basis functions. Standard methods for converging the SCF equations are often found to display divergence when multiple diffuse functions are present, and combinations of steepest descent and second order optimization methods are required. It is furthermore important that the SCF solutions are verified to be genuine minima in the parameter space, as convergence to saddle points is very common when near-degenerate orbitals are present.

In order to check the coherency of the results and detect numerical problems, we have performed parallel calculations with locally modified versions of the Gaussian-09,¹⁷ Gamess-US,¹⁸ and Dalton¹⁹ programs and verified that the results are internally consistent.

Atomic populations have been done using the natural population analysis method.²⁰ The amount of unbound density can be obtained by numerical integration or, more conveniently, by placing the very diffuse basis functions on a dummy atom, in which case the unbound density is simply the electron population for this atom.

III. Results and Discussion

Table 1 shows electron affinities (EA) for first row atoms calculated with four different methods using the aug-pc-2 basis set,¹² which is of triple- ζ quality augmented with diffuse functions, as well as the basis set limiting results obtained as described in more detail in the next section. The pc- n basis sets have been optimized for DFT calculations,²¹ but results similar to those with aug-pc-2 can be obtained using other basis sets of similar quality, as for example the 6-311+G basis set. The four methods are Hartree–Fock (HF)

and three density functional, BHLYP,²² B3LYP³, and BLYP, where the exchange energy is modeled by the Becke gradient corrected functional,²³ and the LYP functional²⁴ is used for calculating the correlation energy. These three functionals can be considered as successively replacing the HF exchange energy by the corresponding DFT analogue. The HF method thus contains 100% HF exchange (and no correlation), BHLYP contains 50% HF and 50% Becke exchange, B3LYP is 20% HF and 80% Becke exchange, while BLYP is 100% Becke exchange.

The results in Table 1 show that the HF method systematically underestimates the EAs due to the lack of electron correlation and only predicts positive EAs for the carbon and fluorine atoms. The DFT methods in general perform much better, and the combination of the B3LYP method and the aug-pc-2 basis set reproduces the six experimentally nonzero EAs (H, Li, B, C, O, F) with an average deviation of only 18 kJ/mol (0.19 eV, 1 eV = 96.5 kJ/mol), and such calibration studies lie at the root of promoting DFT methods with medium basis sets for calculating electron affinities.²⁵

A closer inspection of the results in Table 1, however, reveals a systematic trend. For the HF and BHLYP methods, there are only small changes by extending the size of the basis set beyond aug-pc-2 for atoms predicted to have positive EAs (BHLYP MAD value changes from 19.3 to 18.4 kJ/mol); however, as the amount of HF exchange is reduced, the effect of basis set extension increases to a mean value of 24 kJ/mol for the BLYP method.

The beryllium and nitrogen atoms do not have stable anions and consequently have EAs of zero. In finite basis sets, like aug-pc-2, the extra electron relative to the neutral atom is confined by the basis set to remain close to the nucleus, and the EA calculated as a difference between total energies is often found to be negative. This is a well-known limitation, and calculated negative EAs are therefore usually associated with *de facto* zero EAs. With the aug-pc-2 basis set, all four methods predict a negative EA for the beryllium atom, but the three DFT methods predict a small positive EA at the basis set limit. Somewhat more troublesome is the behavior for the nitrogen atom, where the B3LYP and BLYP methods predict a positive EA even with the aug-pc-2 basis set, and the BLYP value at the basis set limit is a sizable 108 kJ/mol.

The following sections show that these behaviors are due to DFT methods describing anions as having only part of the extra electron bound, while the remaining part drifts off to infinity when a sufficiently flexible basis set is used, where the fraction of the electron bound depends on the magnitude of the EA.

III. a. Atomic Systems. The starting point for establishing the basis set limiting EA values in Table 1 is the aug-pc-*n* basis sets.¹² Only s- and p-functions are required to describe the electron density for the atoms in Table 1, and we have used the aug-pc-*n* (*n* = 0,1,2,3,4) basis sets in their uncontracted forms to avoid possible contraction errors. The aug-pc-4 basis set was systematically extended by adding diffuse s- and p-functions by scaling the outer exponent with a factor of $\sqrt{10}$ until the exponent of the most diffuse function dropped below 10^{-10} . An s- or p-type Gaussian

basis function with an exponent of 10^{-10} has a maximum in the corresponding integrated density corresponding to $\sim 10^5$ au. While basis functions with exponents $\sim 10^{-4}$ are sufficient to describe electrons that are essentially unbound relative to the atom, diffuse functions with exponents $\sim 10^{-8}$ are required to allow the unbound electron density to spread sufficiently to converge the total energy to micro-Hartree accuracy and thus establish the basis set limiting EA.

The hydrogen anion is the simplest system, having only two electrons, and provides a framework for describing the general behavior. In a conventional description, the two electrons are placed in one doubly occupied orbital, which at the HF/aug-pc-2 level of theory leads to an EA of -31.8 kJ/mol. This value changes by only 0.1 kJ/mol when the basis set is extended as described above, i.e. at the basis set limit the HF EA value is -31.7 kJ/mol. The negative value indicates that the total energy of the anion is higher than that of the neutral atom, which is due to the restriction that both electrons are described by the same spatial orbital. When the α and β spin-orbitals are allowed to become different, the energy can be lowered by allowing the β spin-orbital to become as diffuse as the basis set allows. This symmetry breaking leads to an energy lowering of 4.4 kJ/mol with the aug-pc-2 basis set, producing the calculated EA value of -27.4 kJ/mol shown in Table 1. When the basis set is made sufficiently flexible, the β electron drifts off to infinity, and the energy of the anion converges to that of the neutral atom and thus a calculated EA of zero at the basis set limit. These results are shown in Table 2, where the values in parentheses are for the symmetric wave function corresponding to restricting the two electrons to the same spatial orbital. The symmetry breaking is also displayed by the HOMO energies shown in Table 3. In a symmetric description, the HOMO energy is negative (-0.046 au), but in the symmetry-broken solution, the HOMO (β spin-orbital) is positive with the aug-pc-2 basis set and converges to zero from above as the basis set is made sufficiently flexible. The positive HOMO energy and negative EA are thus clear indications that the extra electron is unbound at the HF level, but the use of a limited basis set like aug-pc-2 cannot describe this situation.

With the BLYP functional, the calculated EA with the aug-pc-2 basis set is 82.2 kJ/mol (Table 1), and for this basis set only a symmetric solution corresponding to one doubly occupied orbital can be found, but when a sufficient number of diffuse basis functions are added, both symmetric and symmetry-broken solutions can be found. At the basis set limit, the calculated EAs are 109.1 and 116.4 kJ/mol (Table 1) for the two solutions, respectively, and basis set extensions beyond aug-pc-2 thus changes the EA by 26.9 and 34.1 kJ/mol (Table 2), respectively. The HOMO energy, however, is positive with the aug-pc-2 basis set and converges toward zero for both solutions as the basis set limit is approached (Table 3). The BLYP functional thus predicts a formally unbound electron, despite giving a lower total energy for the anion than for the neutral atom.

This contradiction is resolved by inspecting the HOMO of the symmetric solution, where both the HF and BLYP occupied orbitals display the expected maximum at a distance of ~ 1.2 au from the nucleus which then decays toward zero.

Table 2. Calculated Atomic Properties^a

anion	HF		BHHLYP		B3LYP		BLYP	
	$\Delta EA_{2-\infty}$	D_{unbound}	$\Delta EA_{2-\infty}$	D_{unbound}	$\Delta EA_{2-\infty}$	D_{unbound}	$\Delta EA_{2-\infty}$	D_{unbound}
H ⁻	27.4 (0.1)	1.00 (0)	4.1 (0.5)	0.25 (0.01)	16.9 (9.3)	0.31 (0.18)	34.1 (26.9)	0.37 (0.30)
Li ⁻	10.8 (0.1)	1.00 (0)	2.4 (0.9)	0.21 (0.08)	8.7 (6.5)	0.29 (0.22)	19.4 (17.9)	0.38 (0.35)
Be ⁻	88.3	1.00	75.0	0.72	79.6	0.59	94.8	0.60
B ⁻	26.2 (-0.1)	1.00 (0)	16.1 (5.6)	0.43 (0.17)	27.8	0.41	43.7	0.45
C ⁻	-0.3	0	0.0	0	13.9 (10.2)	0.25 (0.18)	33.0	0.33
N ⁻	160.1	1.00	49.8	0.61	55.8	0.48	73.0	0.48
O ⁻	52.6	1.00	-0.5	0	15.7 (11.7)	0.25 (0.19)	36.8	0.25
F ⁻	-1.3	0	-1.0	0	-0.6	0	10.5 (9.6)	0.15 (0.14)

^a Values correspond to lowest energy solution, which may be symmetry-broken. Values in parentheses are for symmetric solutions, if both types of solutions exist. $\Delta EA_{2-\infty}$ is the change in the calculated electron affinity between the aug-pc-2 basis set and the basis set limit (kJ/mol). D_{unbound} is the amount of unbound electron density in units of one electron when the basis set is made sufficiently flexible.

Table 3. Calculated HOMO Energies (au)^a

anion	HF		BHHLYP		B3LYP		BLYP	
	apc-2	limit	apc-2	limit	apc-2	limit	apc-2	limit
H ⁻	+0.026 (-0.046)	0 (-0.046)	+0.004	0	+0.036	0	+0.065	0
Li ⁻	+0.001 (-0.014)	0 (-0.015)	+0.007	0	+0.022	0	+0.037	0
Be ⁻	+0.033	0	+0.065	0	+0.077	0	+0.087	0
B ⁻	-0.026	0 (-0.027)	+0.023	0	+0.047	0	+0.067	0
C ⁻	-0.078	-0.078	+0.0004	-0.0005	+0.041	0	+0.073	0
N ⁻	+0.056	0	+0.055	0	+0.084	0	+0.110	0
O ⁻	-0.074	0 (-0.074)	-0.005	-0.005	+0.048	0	+0.088	0
F ⁻	-0.181	-0.181	-0.073	-0.073	-0.0006	-0.002	+0.053	0

^a apc-2 indicates results obtained with the aug-pc-2 basis set; limit indicates the basis set limiting results. Values correspond to the lowest energy solution, which may be symmetry-broken. Values in parentheses are for symmetric solutions, if both types of solution exist.

At long distances, however, the BLYP HOMO displays a wavelike behavior with significantly nonzero MO coefficients for the most diffuse basis functions. Figure 1 shows the integrated HF and BLYP HOMO densities normalized to one as a function of the distance from the nucleus, and it is clear that the BLYP method leads to a description where 15% of the density (0.30 electrons) has been expelled from the atom. In a sufficiently flexible basis set, the BLYP method thus describes the H⁻ system as a hydrogen atom with 0.70 extra electrons attached, while the remaining density corresponding to 0.30 electrons behave as free electrons. The BLYP predicted positive EA in a limited basis set like aug-pc-2 is therefore a combination of an energy lowering due to addition of density corresponding to 0.70 electrons, which is partly compensated by an energy increase from constraining a density of 0.30 electrons to occupy the same physical space. In a sufficiently flexible basis set, the latter effect is removed, which accounts for the 26.9 kJ/mol increase in the calculated EA upon enlarging the basis set (Table 2). The symmetry-broken solution provides the same qualitative picture, where a density corresponding to 0.37 (β) electrons becomes unbound, which leads to a change in the EA of 34.1 kJ/mol (Table 2).

The detachment of electrons from the anion can be understood by considering the HOMO and LUMO energies

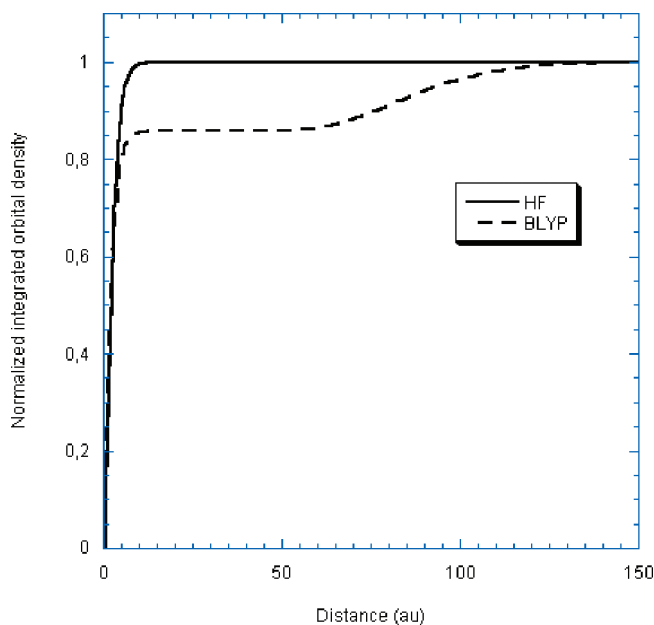


Figure 1. Normalized integrated orbital density for the H⁻ ion as a function of distance from the nucleus using the HF and BLYP methods.

from a SCF calculation with integer electron occupancy. As basis functions with successive smaller exponents are added,

the LUMO energy will converge to zero from above. If the HOMO energy is positive, the LUMO energy will at some stage become lower than the HOMO energy, at which point a lower energy solution can be obtained by transferring electron density from the HOMO to LUMO. The formation of unbound electron density will therefore always occur when a HOMO with positive energy is obtained in a limited basis set and a sufficient number of diffuse basis functions are added. Furthermore, the amount of unbound density will be related to the magnitude of the positive HOMO energy. As electron density is removed from the HOMO, its energy decreases until it becomes degenerate with the formal LUMO. In the limit of a sufficient number of very diffuse basis functions, the LUMO energy will be zero regardless of the amount of electron density it contains, and the unbound electron density is thus the amount required to reduce the HOMO energy to zero.

The B3LYP and BHLYP methods with 20% and 50% HF exchange behave intermediate between the BLYP and HF results, as shown in Tables 2 and 3. Both these hybrid functionals produce positive HOMO energies that converge toward zero as the basis set is augmented with multiple diffuse functions and thus describe partly unbound electron density. With the B3LYP functional, the unbound fraction is 0.31 electrons for the lowest energy symmetry-broken solution, while it is reduced to 0.25 electrons with the BHLYP method. The changes in the calculated EA between the aug-pc-2 and basis set limit are correspondingly smaller than for the BLYP functional. The unbound density using the LSDA functional (not shown) is 0.11 and 0.27 for the symmetric and symmetry-broken solutions, respectively, which is in agreement with the study of Shore et al.⁶

The lithium anion behaves very similar to the hydrogen anion, with slightly larger fractions of electrons being unbound, while the more diffuse character of the 2s-orbital leads to slightly smaller changes in the calculated EA upon improving the aug-pc-2 basis set to the basis set limit, as shown in Table 2.

The carbon atom has a positive EA at the HF level and the anion a negative HOMO energy. With the BHLYP functional the HOMO energy is slightly positive (+0.0004 au) with the aug-pc-2 basis set but becomes negative with the aug-pc-3 basis set and converges to a value of -0.0005 au at the basis set limit (Table 3), while the B3LYP and BLYP functionals produce positive HOMO energies for all basis sets. In a symmetric description, the three singly occupied p-orbitals are equivalent, but when two sets of diffuse functions have been added to the aug-pc-4 basis set, a symmetry-broken description can also be found. In the symmetric solution, the energy of the three degenerate p-orbitals converges toward zero as the basis set limit is approached, while the symmetry-broken solution allows two of the p-orbitals to acquire negative energies, and the third p-orbital has a positive energy that converges toward zero as the basis set limit is attained. With the B3LYP functional, the symmetric solution corresponds to ejecting 0.18 electrons, while the symmetry-broken solution corresponds to having 0.25 electrons unbound. When the BLYP functional is used, only symmetry-broken solutions can be found when more

than three diffuse functions are added to the aug-pc-4 basis set, and this leads to 0.33 electrons being unbound at the basis set limit.

The boron and oxygen anions behave very similar to the carbon anion, with the results shown in Tables 2 and 3.

The fluorine anion has, as mentioned in the Introduction, been used as a test system for probing whether the BLYP functional predicts a bound system. Galbraith and Schaefer calculated a positive HOMO energy of 0.053 au when basis functions with exponents down to $2 \cdot 10^{-5}$ were used,⁹ while Jarecki and Davidson found that extension of the aug-cc-pV5Z basis set with diffuse p-exponents down to $7 \cdot 10^{-9}$ gave a negative HOMO energy of -0.065 au.¹¹ Using the same basis set and grid, we were initially unable to reproduce the latter result, until it was discovered that the behavior could be reproduced by one of the employed programs when the two-electron integrals accuracy was lowered to $\sim 10^{-11}$, which is a typically default setting in many programs. Using our prescription for approaching the basis set limit and employing full integral accuracy, the HOMO energy converges to a value of zero, and the shape of the HOMO leads to a density plot analogous to Figure 1, where the amount of unbound density is 0.14 electrons in a fully symmetric description. It is again possible to generate a symmetry-broken solution when a sufficient number of diffuse functions have been added, and this leads to a marginal increase of the EA by 0.9 kJ/mol and to 0.15 electrons being unbound. The BHLYP and B3LYP functionals have negative HOMO energies when basis sets of aug-pc-2 quality or better is used and thus predicts a completely bound electron (Tables 2 and 3).

The beryllium and nitrogen anions are experimentally found not to be stable (EA = 0 kJ/mol) and are calculated to have positive HOMO energies with all four methods. At the HF level, the extra electron becomes unbound and the EA converges toward zero as the basis set is enlarged. For the beryllium anion, the three DFT methods lead to negative EAs with the aug-pc-2 basis set but positive values at the basis set limit. For the nitrogen anion, the BHLYP calculated EA is negative with the aug-pc-2 basis set, while both the B3LYP and BLYP functionals predict positive EAs. As the basis set is increased toward the limit, all three DFT methods predict positive EAs, and the BLYP value is a sizable 108 kJ/mol (Table 1). In analogy with the other anions, the DFT description is in terms of a fractional EA with binding of 0.28–0.52 electrons (Table 2).

For the atoms with experimentally nonzero EAs, the amount of unbound density in general increases as the amount of HF exchange is reduced. Note that for these systems, the HF-LYP method (100% HF exchange + LYP correlation) predicts a completely bound electron. For the two atoms with zero EA (beryllium and nitrogen), the HF-LYP method predicts a completely unbound electron, and the trend is reversed, such that the amount of unbound density decreases as the amount of HF exchange is reduced (within numerical accuracy).

III. b. Molecular Systems. It has been argued that the problem with positive HOMO energies will be most severe for atoms, where the extra electron is strongly localized, and

Table 4. Calculated Molecular Properties with Four Different Methods^a

anion	HF			BHLYP			B3LYP			BLYP			EA _{Exp}
	ϵ_{HOMO}	EA	D_{unbound}	ϵ_{HOMO}	EA	D_{unbound}	ϵ_{HOMO}	EA	D_{unbound}	ϵ_{HOMO}	EA	D_{unbound}	
CN ⁻	-0.192	280.8	0	-0.106	389.7	0	-0.046	389.8	0	+0.001	361.6 (361.6)	0.003	372.6 ± 0.4
OF ⁻	-0.167	130.2	0	-0.055	233.4	0	+0.019	254.2 (252.7)	0.06	+0.072	256.3 (235.1)	0.22	219.2 ± 0.6
C ₅ O ₂ ⁻	-0.068	60.1	0	+0.003	84.2 (84.1)	0.04	+0.048	92.9 (72.5)	0.32	+0.085	88.3 (40.4)	0.43	116 ± 19
NS ⁻	-0.086	142.8	0	-0.013	152.8	0	+0.037	156.9 (146.3)	0.21	+0.076	153.5 (119.5)	0.33	115.2 ± 1.1
NO ⁻	-0.087	49.5	0	+0.020	84.8 (78.7)	0.23	+0.081	118.9 (75.6)	0.39	+0.121	130.3 (53.5)	0.45	2.5 ± 0.5

^a ϵ_{HOMO} is in au using the aug-pc-2 basis set, EA is in kJ/mol at basis set limit with the aug-pc-2 value in parentheses, and D_{unbound} is in units of one electron.

molecular systems which allow the electron to spread out will be less prone to having unbound electrons.¹⁰ Table 4 shows HF, BHLYP, B3LYP, and BLYP calculated properties for a selection of small molecular systems arranged according to their experimental EA. The HOMO energy is the value obtained with the aug-pc-2 basis set, and for cases where this is positive, calculations have in analogy with the atomic case been performed where multiple diffuse s- and p-functions are added to the aug-pc-2 basis set using a single expansion center. The EAs shown in Table 4 are the values converged with respect to addition of very diffuse functions, with the results for the regular aug-pc-2 basis set shown in parentheses. It was checked that inclusion of very diffuse d-functions had only a marginal influence (less than 0.1 kJ/mol) on the resulting EAs.

The strongly bound CN⁻ anion has a negative HOMO energy with the HF, BHLYP, and B3LYP methods, while the value is slightly positive with the BLYP functional. The amount of unbound electron density in the latter case, however, is only 0.003 electrons, and this has an insignificant effect on the calculated EA, and all three DFT methods provide good estimates of the experimental EA.

The OF⁻ anion represents an intermediate case, where the HF and BHLYP methods give negative HOMO energies, while the B3LYP and BLYP functionals produce positive HOMO energies. The amount of unbound density in the two latter cases is 0.06 and 0.22 electrons, respectively, and for the BLYP functional, the EA changes by 21 kJ/mol when the unbound density is allowed to escape. The values in Table 4 correspond to a symmetric solution where the α and β spin-orbitals are constrained to be identical, but in analogy with the atomic systems, it is possible to generate symmetry-broken solutions where the α and β spin-orbitals are inequivalent. The symmetry-broken solution further increases the EA by 2 kJ/mol and corresponds to 0.24 electrons being unbound.

The NS⁻ and NO⁻ anions are triplet ground states and have smaller EAs, with the latter being near-zero experimentally. The HF method predicts negative HOMO energies for both species, while the B3LYP and BLYP methods have positive HOMO energies, and the BHLYP method predicts a positive HOMO energy for NO⁻ and a negative HOMO energy for NS⁻. For the cases where the HOMO energy is positive only symmetry-broken solutions, where the π_x - and π_y -orbitals are inequivalent, could be found when a sufficient

number of diffuse functions were added. The NS⁻ and NO⁻ results follow the trend that the amount of unbound density increases as the EA becomes lower and increases as the amount of HF exchange is decreased. With the BLYP functional, the NS⁻ and NO⁻ systems have 0.33 and 0.45 electrons unbound, respectively, and the changes in the calculated EA upon addition of multiple diffuse functions are 34 and 77 kJ/mol. The B3LYP and BHLYP results are again intermediate between the BLYP and HF results.

The C₅O₂⁻ anion, which has an experimental EA very similar to NS⁻, albeit with a rather large uncertainty, also displays the phenomenon of fractional EA with all three DFT methods (Table 4). The C₅O₂⁻ system has slightly more positive HOMO energies and slightly more unbound density than NS⁻, and the calculated EAs are actually closer to that of NO. The similarity of the C₅O₂⁻ results with those of NS⁻ and NO⁻ suggest that delocalization of the extra electron over a few atoms is not enough to overcome the inherent tendency for a given XC functional to predict unbound electron density. As the system size increases, the delocalization effect will of course at some point become important, and for example the C₆₀⁻ anion is calculated to have a negative HOMO energy with the BLYP functional. Note, however, that the delocalization effect results in the HF method predicting positive EAs for all five systems. The results in Table 4 indicate that many commonly encountered anions will display the phenomenon of fractional EA when using standard DFT methods.

III. c. Intermolecular Complexes, Zwitterions, and Donor–Acceptor Systems. The DFT description of anions as having fractional unbound electron density has consequences for systems where (localized) anions are part of a larger system that also have moieties that can accept electron density, even when using regular basis sets and grids. To illustrate this point, we have performed BLYP/6-31+G* calculations⁴ for the intermolecular complex of a formate anion (HCO₂⁻) and an ammonium cation (NH₄⁺) as a function of the distance between the carbon and nitrogen atoms, as shown in Figure 2. At the HF and MP2 levels of theory using the restriction of doubly occupied orbitals, the complex separates into two ions having a full negative and positive charge, and this charge separation is essentially complete for distances longer than 6 Å, as shown in Figure 3. At large separations, a similar solution can be found also with the BLYP functional; however, this is a saddle point

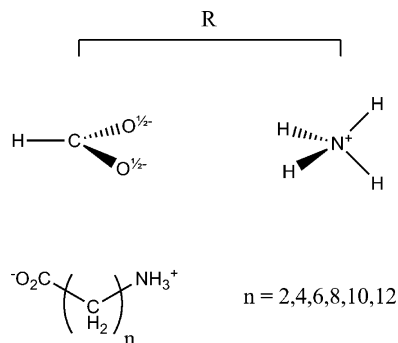


Figure 2. Molecular complex and zwitterions used for the results in Figures 3 and 4.

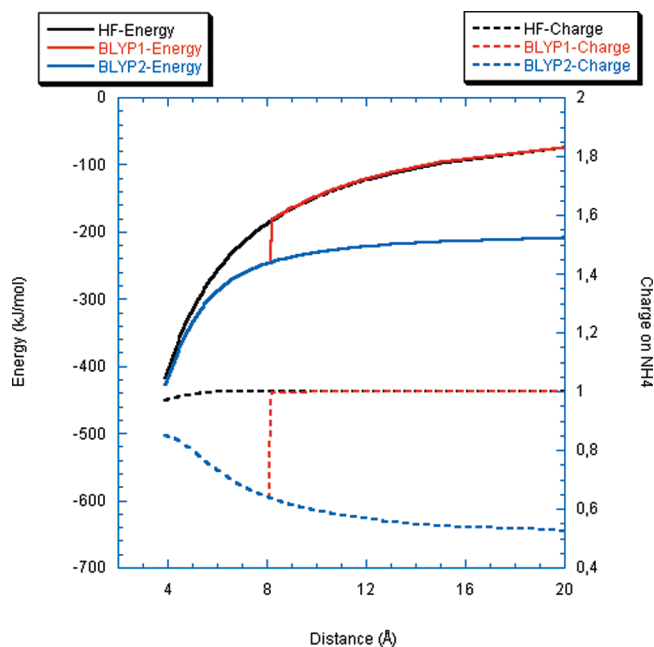


Figure 3. Energies relative to the ion separated limit and charge on the NH_4 group for the $\text{HCO}_2^- - \text{NH}_4$ complex in Figure 2 as a function of separation distance. BLYP1 and BLYP2 indicate two different SCF solutions, see text for details.

in the parameter space. The lowest energy solution corresponds to the formate ion having transferred 0.53 electrons to the ammonium group, and at the dissociation limit this is 189 kJ/mol lower in energy than the complete charge-separated solution. For distances smaller than 8.2 Å, only a single SCF solution can be found, and from the continuity of the energy and charge curves in Figure 3, the solution at shorter distances is of the fractionally transferred electron type. It should be noted that several other solutions corresponding to intermediate fractional electron transfer can be found in the medium distance range of 9–13 Å, but these have been omitted in Figure 3 for clarity.

The two other XC functionals display a behavior intermediate between the BLYP and HF results. The popular B3LYP method thus dissociates into a formate ion having transferred 0.45 electrons to the ammonium group, leading to an energy lowering of 115 kJ/mol relative to the completely ion-separated description, while the corresponding values for the BHLYP method are 0.27 electrons and 35 kJ/mol. The collapse of the ionic solution occurs at a

distance of 7.0 Å with the B3LYP method and 9.0 Å with the BHLYP functional.

At large separation distances, the employed programs converge to the pure ionic solution when using automated start guesses for the SCF procedure, and manual procedures must be employed to locate the fractional electron transfer solution. At intermediate distances, the convergence using the automated start guesses is erratic, and which SCF solution that is obtained depends on numerical details.

When using an unrestricted wave function, the HF solution corresponding to dissociation into neutral formate and ammonium radicals is 13 kJ/mol higher in energy than the ion-pair dissociation, while the MP2 level favors the neutral dissociation by 19 kJ/mol, and the energy difference can be taken as the difference in EA between the formate radical and the ammonium cation. The three DFT methods also predict a neutral dissociation channel to be lowest, by 9, 51, and 88 kJ/mol for the BHLYP, B3LYP, and BLYP methods, respectively. It may be argued that the DFT results with fractional electron transfer are trying to mimic this radical dissociation within a restricted framework using doubly occupied orbitals. The fact that the fraction of density transferred dependent on the XC functional, that the BLYP energy difference between the two solutions is 189 kJ/mol, compared to the near zero values at the HF and MP2 levels, and the similarity with the ejection of density from the isolated anions, suggest that fractional electron transfer is an artifact of the DFT methods, rather than an attempt of describing the actual physical system.

The formate ammonium complex in Figure 2 can be taken as a model for negatively and positively charged side chains in peptides and proteins. If these are in close contact and form a salt bridge, Figure 3 suggests that DFT methods will predict too little charge separation between the two moieties due to fractional electron transfer. Furthermore, they will display an incorrect behavior for how the atomic charges change as a function of separation distance, i.e. Figure 3 shows that the charge separation becomes smaller as the distance is increased instead of becoming larger. If a number of anionic and cationic sites are separated by more than a few Ångströms, there may be multiple SCF solutions corresponding to combinations of fractional electron transfer, and which solution a given program finds may depend on numerical details.

The DFT description of fractional electron transfer is not limited to intermolecular ion complexes but is also present in zwitterionic systems, where the anionic and cationic sites are part of the same molecule. For the ammonium-alkyl-carboxylates shown in Figure 2, the charges on the NH_3 and CO_2 groups are shown in Figure 4 as a function of chain length ($n = 2, 4, 6, 8, 10, 12$) at the HF and BLYP levels of theory with the 6-31+G* basis set. For the two longest chains, there are two BLYP solutions, where the one with the lowest energy is of the fractional electron transfer type. For $n = 12$, corresponding to a distance between the terminal carbon and nitrogen atoms of 16.6 Å, the energy difference is 86 kJ/mol and has ~ 0.31 electrons transferred from the carboxylate to the ammonium group. For the systems with eight or less methylene groups, only a single BLYP solution

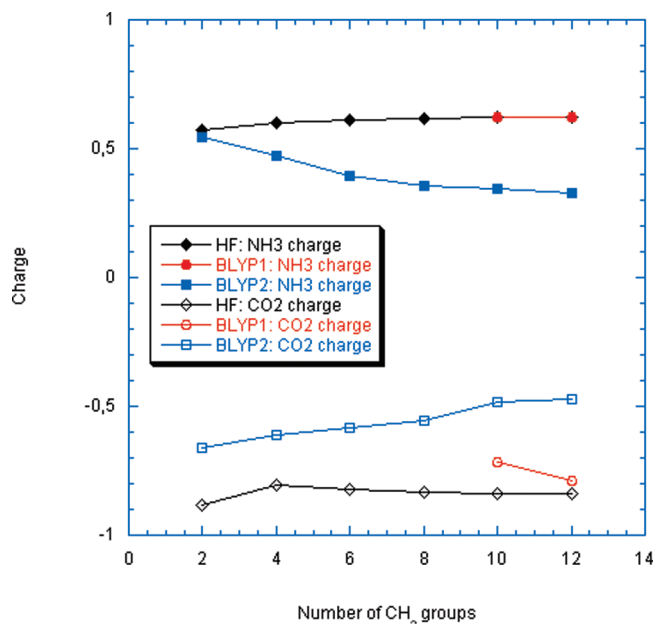


Figure 4. Charge on the CO₂ and NH₃ groups for the zwitterionic systems in Figure 2 as a function of number of methylene units. BLYP1 and BLYP2 indicate two different SCF solutions, see text for details.

can be found, and from the continuity of the curves in Figure 4, this is of the fractional electron transfer type. The B3LYP method produces results intermediate between the HF and BLYP, with two different solutions for the two longest chains, where ~ 0.22 electrons have been transferred in the lowest energy solution.

The fractional electron transfer displayed by large distance separated ionic system, being either inter- or intramolecular, is a direct consequence of the fractional unbound electron density for isolated anions discussed in the previous sections. In the general case, however, it is likely that the phenomenon will be present when a sufficiently electron-rich donor is present together with a sufficiently electron-poor acceptor, where neither of these needs to carry a full negative or positive charge. To illustrate this point, we have carried out calculations of two systems derived from the formate and ammonium ions in Figure 2. If the formate anion is protonated to give a formic acid-ammonium complex, the system still displays fractional electron transfer with the BLYP functional, although only 0.02 electrons are transferred to the ammonium group. Similarly, the formate anion complexed with neutral ammonia leads to transfer of 0.02 electrons in the large separation limit. The energy differences between the pure ionic and fractional electron transfer solutions are only a few tenths of a kJ/mol. With the B3LYP functional, both of these systems have completely localized charges, again confirming that pure density functional methods are most prone to fractional electron transfer. The small amounts of electron transfer with the BLYP functional will only have notable effects for intermolecular complexes separated by a large distance and will most likely only have small consequences for intramolecular systems. It is likely, however, that there is a continuous range of fractional electron transfer, when the amount of density transferred depends on the strengths of the electron donor and acceptor,

with isolated anions and cations representing the limiting case, and the XC functional. How severe this problem is will have to be settled by more detailed investigations for a larger variety of systems.

The results in Figures 3 and 4 suggest that standard DFT methods will be problematic for describing the electronic structure of anions and cations embedded in a large macromolecular system, like a protein. Assuming that the employed program is capable of locating the lowest energy SCF solution, the corresponding description will have too little charge separation and leads to incorrect descriptions of the electrostatic potential.

III. d. Dependence on XC Functional. We have in the above analysis shown results for the BHLYP, B3LYP, and BLYP functionals, as these represent some of the most widely used methods. As shown by the results in Tables 2–4, the presence of a positive HOMO energy indicates that the method will display the phenomenon of fractional EA, and this may have significant consequences when HOMO energies calculated with a diffuse augmented basis set of double or triple- ζ quality are larger than ~ 0.01 au. There is a clear, although not quantitative, correlation between the magnitude of the EA and the HOMO energy, i.e. species with large EAs have negative or small positive HOMO energies, while species with low EAs have positive HOMO energies with all three XC functionals. Fractional electron transfer in inter- or intramolecular complexes can occur even in the absence of positive HOMO energies but can be detected by the presence of at least one virtual orbital with lower energy than the HOMO. Allowing these orbitals to mix will produce degenerate HOMO and LUMO orbitals with fractional occupation.

We have scanned a large part of the XC functionals available in the employed computational packages^{17,18} for the B⁻, C⁻, O⁻, F⁻, CN⁻, OF⁻, NS⁻, and NO⁻ anions using the aug-pc-2 basis set and find that the results are representative for all of the commonly employed functionals, including meta-functionals like TPSS and TPSSH.²⁶ Pure XC functionals that do not contain HF exchange resemble the BLYP results in having the largest positive HOMO energies, while hybrid methods like B3LYP and BHLYP have negative HOMO energies for the systems with large EAs but positive energies for systems with small EAs. The larger the amount of HF exchange in the functional, the lower HOMO energies are obtained, and functionals that employ 100% HF exchange display negative HOMO energies for all the tested species.

The dependence on the amount of HF exchange present in the XC functional clearly suggests that it is the incorrect distance dependence of the exchange functional²⁷ that is the reason for the fractional electron affinity phenomenon. In order to test this hypothesis, we have performed calculations with the long-range corrected version of the BLYP functional (LC-BLYP), where the amount of HF exchange increases continuously as a function of electron–electron separation distance,²⁸ and this indeed solves the problem. All the species with positive EAs are calculated to have negative HOMO energies for the corresponding anion with this functional, and no fractional electron transfer for the intermolecular complex or zwitterionic systems shown in Figure 2 is

observed. The LC-BLYP method correctly predicts a zero EA for the beryllium atom, predicts a small nonzero EA for the nitrogen atom (6 kJ/mol), and significantly overestimates that EA for NO by 91 kJ/mol but at least provides a qualitative correct description as having either zero or one electron bound.

IV. Summary

It is shown that the most commonly used DFT methods describe anions as having only a fraction of the extra electron bound, where the amount of bound electron density depends on the magnitude of the EA and the XC functional. Pure functionals without HF exchange display the largest effect, while hybrid methods with increasingly larger amounts of HF exchange behave intermediate between the pure functional and the pure HF methods. For species with large EAs, the fraction of electron bound is close to or equal to one, even for pure functionals like BLYP, and DFT methods provide estimates of the experimental EA within the expected DFT accuracy. For species with small EAs, however, most commonly employed functionals predict a significant overbinding when very large basis sets are used. When using standard basis sets, the unbound density is constrained to remain close to the nucleus, and the resulting Coulomb repulsion compensates partly for the inherent overbinding, which explains the observed reasonable correlation between DFT calculated and experimental EAs.

For inter- and intramolecular systems having both anionic and cationic sites, the phenomenon of fractional electron affinity results in fractional electron transfer SCF solutions even for regular basis sets. At large separation these solutions have a much lower energy than the ionic separation, and only the fractional electron transfer solution is present at short separation distances. This phenomenon is likely to be present in any system containing sufficiently electron-rich donors and electron-poor acceptors, at least for pure XC functionals. Clearly the existence of such fractional electron transfer descriptions will have consequences for the use of DFT methods to model ionic and polar systems.

Establishing the amount of unbound density for an isolated anion is technically somewhat difficult, as few programs are set up to handle very diffuse basis functions. Fortunately, the presence of a positive HOMO energy with a medium sized basis set is a clear indicator of this situation^{29,30} and can be used as a diagnostics for whether a given XC functional can describe a given anion. Similarly, if a virtual orbital is lower in energy than the HOMO energy, this indicates that a lower energy SCF solution exists, and if the two orbitals are spatially well separated, this will be of the fractionally electron transfer type.

The origin of the phenomenon of fractional electron affinity is the incorrect long-range behavior of the exchange functional, and a possible solution is to apply the long-range correction method to the exchange functional. Using such long-range corrected functionals is strongly recommended when using density functional methods to describe the electronic structure of systems with localized anionic or strongly electron donating sites.

After submission of the present paper, Lee et al. published a paper with a detailed analysis of the XC potential for the lithium anion and arrived at conclusions in agreement with those in the present paper.³⁰

Acknowledgment. This work was supported by grants from the Danish Center for Scientific Computation and the Danish Natural Science Research Council.

References

- (1) Koch, W.; Holthausen, M. C. *A Chemist's Guide to Density Functional Theory*; Wiley-VCH: 2001. (b) Jensen, F. *Introduction to Computational Chemistry*; Wiley: 2006.
- (2) Hohenberg, P.; Kohn, W. *Phys. Rev.* **1964**, *136*, B864.
- (3) (a) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648. (b) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. *J. Phys. Chem.* **1994**, *98*, 11623.
- (4) (a) Hehre, W. J.; Ditchfield, R.; Pople, J. A. *J. Chem. Phys.* **1972**, *56*, 2257. (b) Frisch, M. J.; Pople, J. A.; Binkley, J. S. *J. Chem. Phys.* **1984**, *80*, 3265. (c) Francl, M. M.; Pietro, W. J.; Hehre, W. J.; Binkley, J. S.; Gordon, M. S.; DeFrees, D. J.; Pople, J. A. *J. Chem. Phys.* **1982**, *77*, 3634.
- (5) (a) de Oliveira, G.; Martin, J. M. L.; de Proft, F.; Geerling, P. *Phys. Rev. A* **1999**, *60*, 1034. (b) Boese, A. D.; Oren, M.; Atasoylu, O.; Martin, J. M. L.; Kállay, M.; Gauss, J. *J. Chem. Phys.* **2004**, *120*, 4129.
- (6) Shore, H. B.; Rose, J. H.; Zaremba, E. *Phys. Rev. B* **1977**, *15*, 2858.
- (7) (a) Schawrz, K. *Chem. Phys. Lett.* **1978**, *57*, 605. (b) Sen, K. D. *Chem. Phys. Lett.* **1980**, *74*, 201. (c) Cole, L. A.; Perdew, J. P. *Phys. Rev. A* **1982**, *25*, 1265. (d) Guo, Y.; Whitehead, M. A. *Phys. Rev. A* **1989**, *40*, 28.
- (8) Vydrov, O. A.; Scuseria, G. E. *J. Chem. Phys.* **2005**, *122*, 184107.
- (9) Galbraith, J. M.; Schaefer, H. F., III. *J. Chem. Phys.* **1996**, *105*, 862.
- (10) Rösch, N.; Tricky, S. B. *J. Chem. Phys.* **1996**, *106*, 8940.
- (11) Jaręcki, A. A.; Davidson, E. R. *Chem. Phys. Lett.* **1999**, *300*, 44.
- (12) Jensen, F. *J. Chem. Phys.* **2002**, *117*, 9234.
- (13) Rienstra-Kiracofe, J. C.; Tschumper, G. S.; Schaefer, H. F., III; Nandi, S.; Ellison, G. B. *Chem. Rev.* **2002**, *102*, 231.
- (14) Some recent examples: (a) Chattarai, P. K.; Duley, S. *J. Chem. Eng. Data* **2010**, *55*, 1882. (b) Karwowski, B. T. *Cent. Eur. J. Chem.* **2010**, *8*, 70. (c) Gong, L. F.; Wu, X. M.; Li, W.; Qi, C. S.; Xiong, J. M.; Guo, W. L. *Mol. Phys.* **2009**, *107*, 701. (d) Feng, X. J.; Li, Q. S.; Gu, J. D.; Cotton, F. A.; Xie, Y. M.; Schaefer, H. F., III. *J. Phys. Chem. A* **2009**, *113*, 887.
- (15) Murray, C. W.; Handy, N. C.; Laming, G. J. *Mol. Phys.* **1993**, *78*, 997.
- (16) Lebedev, V. I.; Laikov, D. N. *Doklady Math.* **1999**, *59*, 477.
- (17) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Veven, T.; Montgomery, Jr., J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark,

- M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, O.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. *Gaussian 09*; Gaussian, Inc.: Wallingford, CT, 2009.
- (18) Schmidt, M. W.; Baldrige, K. K.; Boatz, J. A.; Elbert, S. T.; Gordon, M. S.; Jensen, J. J.; Koseki, S.; Matsunaga, N.; Nguyen, K. A.; Su, S.; Windus, T. L.; Dupuis, M.; Montgomery, J. A. *J. Comput. Chem.* **1993**, *14*, 1347.
- (19) Angeli, C.; Bak, K. L.; Bakken, V.; Christiansen, O.; Cimiraglia, R.; Coriani, S.; Dahle, P.; Dalskov, E. K.; Enevoldsen, T.; Fernandez, B. Hättig, C.; Hald, K.; Halkier, A.; Heiberg, H.; Helgaker, T.; Hettema, H.; Jensen, H. J. Aa.; Jonsson, D.; Jørgensen, P.; Kirpekar, S.; Klopper, W.; Kobayashi, R.; Koch, H.; Ligabue, A.; Lutnæs, O. B.; Mikkelsen, K. V.; Norman, P.; Olsen, J.; Packer, M. J.; Pedersen, T. B.; Rinkevicius, Z.; Rudberg, E.; Ruden, T. A.; Ruud, K.; Salek, P.; Sanchez de Meras, A.; Saue, T.; Sauer, S. P. A.; Schimmelpfennig, B.; Sylvester-Hvid, K. O.; Taylor, P. R.; Vahtras, O.; Wilson, D. J.; Ågren, H. <http://www.kjemi.uio.no/software/dalton/dalton.html> (accessed May 6, 2009).
- (20) Reed, A. E.; Curtiss, L. A.; Weinhold, F. *Chem. Rev.* **1988**, *88*, 899.
- (21) (a) Jensen, F. *J. Chem. Phys.* **2001**, *115*, 9113. **2001**, *116*, 3502. (b) Jensen, F. *J. Chem. Phys.* **2001**, *116*, 7372. (c) Jensen, F.; Helgaker, T. *J. Chem. Phys.* **2004**, *121*, 3463.
- (22) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 1372.
- (23) Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098.
- (24) (a) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785. (b) Miehlich, B.; Savin, A.; Stoll, H.; Preuss, H. *Chem. Phys. Lett.* **1989**, *157*, 200.
- (25) (a) Tschumper, G. S.; Schaefer, H. F., III. *J. Chem. Phys.* **1997**, *107*, 2529. (b) Vera, D. M. A.; Pierini, A. B. *Phys. Chem. Chem. Phys.* **2004**, *6*, 2899. (c) Li, X.; Cai, Z.; Sevilla, M. D. *J. Phys. Chem. A* **2002**, *106*, 1596. (d) Puiatti, M.; Vera, D. M. A.; Pierini, A. B. *Phys. Chem. Chem. Phys.* **2008**, *10*, 1394. (e) Papajak, E.; Truhlar, D. G. *J. Chem. Theory Comput.* **2010**, *6*, 597. (f) Knoll, E. H.; Friesner, R. A. *Phys. Chem. B* **2006**, *110*, 18787.
- (26) (a) Tao, J.; Perdew, J. P.; Staroverov, V. N.; Scuseria, G. E. *Phys. Rev. Lett.* **2003**, *91*146401. (b) Staroverov, V. N.; Scuseria, G. E.; Tao, J.; Perdew, J. P. *J. Chem. Phys.* **2003**119, 12129.
- (27) Wu, Q.; Ayers, P. W.; Yang, W. *J. Chem. Phys.* **2003**, *119*, 2978.
- (28) (a) Iikura, H.; Tsuneda, T.; Yanai, T.; Hirao, K. *J. Chem. Phys.* **2001**, *115*, 3540. (b) Yanai, T.; Tew, D. P.; Handy, N. C. *Chem. Phys. Lett.* **2004**, *393*, 51.
- (29) A positive HOMO energy when using the conventional choice of a vanishing Kohn-Sham potential at large distance.
- (30) Lee, D.; Furche, F.; Burke, K. *Phys. Chem. Lett.* **2010**, *1*, 2124.

CT1003324

Electron Localization Function at the Correlated Level: A Natural Orbital Formulation

Ferran Feixas,[†] Eduard Matito,^{*,‡} Miquel Duran,[†] Miquel Solà,[†] and Bernard Silvi^{*,§}

Institut de Química Computacional i Departament de Química, Universitat de Girona, Campus de Montilivi, 17071 Girona, Catalonia, Spain, Institute of Physics, University of Szczecin, Wielkopolska 15, 70-451 Szczecin, Poland, and Laboratoire de Chimie Théorique Université Pierre et Marie Curie 3, rue Galilée 94200 Ivry sur Seine, Paris, France

Received June 26, 2010

Abstract: In this work we present a 2-fold approximation for the calculation of the electron localization function (ELF) which avoids the use of the two-particle density (2-PD). The first approximation is used for the calculation of the ELF itself and the second one is used to approximate pair populations integrated in the ELF basins. Both approximations only need the natural orbitals and their occupancies, which are available for most methods used in electronic structure calculations. In this way, methods such as CCSD and MP2 can be used for the calculation of the ELF despite the lack of a pertinent definition of the 2-PD. By avoiding the calculation of the 2-PD, the present formulation provides the means for routine calculations of the ELF in medium-size molecules with correlated methods. The performance of this approximation is shown in a number of examples.

1. Introduction

In a recent work,¹ we described in detail a correlated version of the Electron Localization Function (ELF) initially proposed by Becke and Edgecombe² for Hartree–Fock (HF) wave functions. This correlated version, derived from the analysis of the pair functions performed independently by one of us³ and by Kohout and co-workers,^{4–7} has been implemented in the TopMoD package.⁸ The correlated ELF requires the calculation of the laplacian of the same-spin pair functions, i.e.,

$$\nabla_{\mathbf{r}_2}^2 (\pi^{\alpha\alpha}(\mathbf{r}_1, \mathbf{r}_2) + \pi^{\beta\beta}(\mathbf{r}_1, \mathbf{r}_2))|_{\mathbf{r}_2=\mathbf{r}_1} \quad (1)$$

while the covariance analysis of the electron population implies the evaluation of the integrated pair densities $\bar{N}^{\alpha\alpha}(\Omega)$, $\bar{N}^{\alpha\beta}(\Omega)$, and $\bar{N}^{\beta\beta}(\Omega)$. Unfortunately, the numerical complexity of the exact approach, which considers the correlated two-particle density (2-PD), limits the use of the correlated ELF population analysis to small systems. Several applications of the ELF include analysis of organometallic complexes,⁹ aromaticity analysis,^{10–12} electronic structure studies along the IRC,^{13,14} and

the mechanism analysis in electrocyclic reactions.^{15,16} Such analyses involve large molecules for which the calculation of the 2-PD is beyond reasonable cost.

It is possible to lower the numerical complexity of the calculation of the ELF as well as that of the pair populations by expressing the 2-PD in terms of natural geminals.^{17,18} This strategy is, however, rather difficult to implement because the determination of the natural geminals requires the full diagonalization of a very large matrix. Approximate expressions of the 2-PD using the first-order reduced density matrix (1-RDM) or the natural spin orbitals have recently been derived in the framework of the density matrix functional theory.^{19–25} The aim of the present work is to investigate the ability of these latter expressions to calculate reliable values of the correlated localization functions and of the integrated pair density.

2. Natural Spin-Orbital Expression of the Reduced Two-Particle Density

The 2-PD is defined as follows:²⁶

$$\pi(\mathbf{r}_1, \mathbf{r}_2) = \int \Psi(x_1, x_2, x_3, \dots, x_N) \Psi^*(x_1, x_2, x_3, \dots, x_N) dx'' d\sigma_1 d\sigma_2 \quad (2)$$

* To whom correspondence should be addressed: E-mail: ematito@gmail.com; silvi@lct.jussieu.fr.

[†] University of Girona.

[‡] University of Szczecin.

[§] Université Pierre et Marie Curie.

$$= \langle \Psi | \hat{\pi}(\mathbf{r}_1, \mathbf{r}_2) | \Psi \rangle \quad (3)$$

$$= \pi^{\alpha\alpha}(\mathbf{r}_1, \mathbf{r}_2) + \pi^{\alpha\beta}(\mathbf{r}_1, \mathbf{r}_2) + \pi^{\beta\alpha}(\mathbf{r}_1, \mathbf{r}_2) + \pi^{\beta\beta}(\mathbf{r}_1, \mathbf{r}_2) \quad (4)$$

In eq 2 dx'' indicates that the integration is performed over the space and spin coordinates of all electrons but two. The 2-PD operator appearing in eq 3 is as follows:

$$\hat{\pi}(\mathbf{r}_1, \mathbf{r}_2) = \sum_{i=1}^N \sum_{j \neq i}^N \delta(\mathbf{r}_i - \mathbf{r}_1) \delta(\mathbf{r}_j - \mathbf{r}_2) \quad (5)$$

Finally in eq 4, $\pi^{\sigma\sigma}(\mathbf{r}_1, \mathbf{r}_2)$ and $\pi^{\sigma\sigma'}(\mathbf{r}_1, \mathbf{r}_2)$ represent the probability densities of finding one electron of spin σ at \mathbf{r}_1 and another at \mathbf{r}_2 , this latter of spin σ and σ' , respectively, regardless the position of the other $N - 2$ electrons.²⁶ The calculation of the ELF expression only needs of the same-spin 2-PD, $\pi^{\sigma\sigma}(\mathbf{r}_1, \mathbf{r}_2)$. The Pauli principle prescribes the following:

$$\pi^{\sigma\sigma}(\mathbf{r}_1, \mathbf{r}_1) = 0 \quad (6)$$

and with the electron–electron cusp condition, it provides the following:²⁷

$$\nabla_{\mathbf{r}_2} \pi^{\sigma\sigma}(\mathbf{r}_1, \mathbf{r}_2) |_{\mathbf{r}_2=\mathbf{r}_1} = 0 \quad (7)$$

whereas $\pi^{\sigma\sigma'}(\mathbf{r}_1, \mathbf{r}_1)$ is usually greater than zero.²⁸ The original definition of the ELF by Becke² assumed eq 6, while a recent formulation,³ which takes the first-non vanishing term of Taylor expansion of the 2-PD assumes both eqs 6 and 7 (it is worth noting that proper choice of the reference point or spherical averaging immediately provides eq 7). In exact wave functions, both conditions come from the antisymmetry of the wave function and are fulfilled by the *exact* 2-PD.²⁷ Therefore, we also need the approximate expression of the 2-PD used in the calculation of the ELF to fulfill these properties.

The integrated pair densities over the whole space should yield the following:

$$\bar{N}^{\sigma\sigma} = \iint \pi^{\sigma\sigma}(\mathbf{r}_1, \mathbf{r}_2) d\mathbf{r}_1 d\mathbf{r}_2 = \bar{N}^{\sigma}(\bar{N}^{\sigma} - 1) \quad (8)$$

$$\bar{N}^{\sigma\sigma'} = \iint \pi^{\sigma\sigma'}(\mathbf{r}_1, \mathbf{r}_2) d\mathbf{r}_1 d\mathbf{r}_2 = \bar{N}^{\sigma} \bar{N}^{\sigma'} \quad (9)$$

These pair densities are used to calculate the covariance matrix of the electron population in the ELF basins. If we use an approximate expression of the 2-PD to calculate these quantities, then it should fulfill these sum rules in order to obtain a meaningful distribution of the N electrons among the ELF basins.

For a linear expansion of the wave function, the 2-PD distribution can be expressed on the basis of the molecular orbitals:

$$\pi(\mathbf{r}_1, \mathbf{r}_2) = \sum_{ijkl} \Gamma_{ij}^{kl} \varphi_i^*(\mathbf{r}_1) \varphi_j^*(\mathbf{r}_2) \varphi_k(\mathbf{r}_1) \varphi_l(\mathbf{r}_2) \quad (10)$$

where the Γ_{ij}^{kl} coefficients obey symmetry and antisymmetry relationships:

$$\Gamma_{ij}^{kl} = \Gamma_{ji}^{lk} = \Gamma_{kl}^{ij} = \Gamma_{lk}^{ji} \quad (11)$$

$$= -\Gamma_{ij}^{lk} = -\Gamma_{ji}^{kl} = -\Gamma_{kl}^{ji} = -\Gamma_{lk}^{ij} \quad (12)$$

These relationships are independent of the type of spinorbitals (canonical or not, provided they are orthonormal) and, in particular, they are valid for natural spinorbitals. Taking advantage of the symmetry and antisymmetry of the coefficients, eq 10 can be rewritten as follows:

$$\pi(\mathbf{r}_1, \mathbf{r}_2) = 2 \sum_{ijkl} \Gamma_{ij}^{kl} (\varphi_i^*(\mathbf{r}_1) \varphi_j^*(\mathbf{r}_2) \varphi_k(\mathbf{r}_1) \varphi_l(\mathbf{r}_2) - \varphi_i^*(\mathbf{r}_1) \varphi_j^*(\mathbf{r}_2) \varphi_l(\mathbf{r}_1) \varphi_k(\mathbf{r}_2)) \quad (13)$$

where the prime indicates that the symmetry relations have been implicitly taken into account in the coefficients in order to restrict the sum to the independent ones. However, insofar as eq 3 is used to evaluate the pair functions, it is not advantageous to use natural orbitals instead of the canonical ones and an approximate expression of $\pi(\mathbf{r}_1, \mathbf{r}_2)$ is highly desirable in order to reduce its computational cost.

Such approximations are derived from a general expression of $\pi(\mathbf{r}_1, \mathbf{r}_2)$ in terms of the 1-RDM elements $\rho(\mathbf{r}_1, \mathbf{r}_2)$, i.e.,

$$\pi(\mathbf{r}_1, \mathbf{r}_2) = \rho(\mathbf{r}_1) \rho(\mathbf{r}_2) - \rho(\mathbf{r}_1, \mathbf{r}_2) \rho(\mathbf{r}_2, \mathbf{r}_1) + \lambda_2(\mathbf{r}_1, \mathbf{r}_2) \quad (14)$$

According to Kutzelnigg and Mukherjee,²¹ the three contributions appearing in eq 14 respectively correspond to the full Coulomb interaction, the exchange interaction, and the correlation correction (the latter two are usually called exchange-correlation contribution, which we shall denote by XC). The natural orbital expressions for the Coulomb and exchange interaction terms involve sums over only two indices, whereas the exact correlation contribution is much more complicated. This latter is however expected to be small and therefore several approximate expressions in terms of the natural orbitals φ_i and their occupancies n_i have been proposed.

1. The HF like approximation for the exchange-correlation part (HF-XC hereafter) assumes,

$$\lambda_2(\mathbf{r}_1, \mathbf{r}_2) = 0 \quad (15)$$

which yields the following NO expression of the 2-PD,

$$\pi(\mathbf{r}_1, \mathbf{r}_2) = \sum_i \sum_j n_i n_j (\varphi_i^*(\mathbf{r}_1) \varphi_j^*(\mathbf{r}_2) \varphi_i(\mathbf{r}_1) \varphi_j(\mathbf{r}_2) - \varphi_i^*(\mathbf{r}_1) \varphi_j^*(\mathbf{r}_2) \varphi_j(\mathbf{r}_1) \varphi_i(\mathbf{r}_2)) \quad (16)$$

This approximation obviously satisfies the symmetry and antisymmetry relationships between coefficients and fulfills the requirements of eqs 6 and 7. However, it violates the sum rule (eq 8) because the HF-like exchange-correlation part (second term in the rhs of eq 14) only integrates to $-N^{\sigma}$ for monodeterminantal wave functions.

2. Goedecker and Umrigar²⁰ have proposed an expression (GU functional):

$$\pi(\mathbf{r}_1, \mathbf{r}_2) = \sum_i \sum_{j \neq i} n_i n_j \varphi_i^*(\mathbf{r}_1) \varphi_j^*(\mathbf{r}_2) \varphi_i(\mathbf{r}_1) \varphi_j(\mathbf{r}_2) - \sqrt{n_i n_j} \varphi_i^*(\mathbf{r}_1) \varphi_j^*(\mathbf{r}_2) \varphi_j(\mathbf{r}_1) \varphi_i(\mathbf{r}_2) \quad (17)$$

which satisfies neither the antisymmetry requirements nor the sum rules. It corresponds to a correlation correction of the form,

$$\lambda_2(\mathbf{r}_1, \mathbf{r}_2) = \sum_i \sum_{j \neq i} (n_i n_j - \sqrt{n_i n_j}) \varphi_i^*(\mathbf{r}_1) \varphi_j^*(\mathbf{r}_2) \varphi_j(\mathbf{r}_1) \varphi_i(\mathbf{r}_2) \quad (18)$$

3. Buijse and Baerends's first approximation^{22,25} (BB functional), which was previously derived by Müller,¹⁹ removes the restriction in the inner sum of the GU functional in order to preserve the sum rules.

$$\lambda_2(\mathbf{r}_1, \mathbf{r}_2) = \sum_i \sum_{j \neq i} (n_i n_j - \sqrt{n_i n_j}) \varphi_i^*(\mathbf{r}_1) \varphi_j^*(\mathbf{r}_2) \varphi_j(\mathbf{r}_1) \varphi_i(\mathbf{r}_2) \quad (19)$$

Holas²⁹ has recently generalized the GU functional, giving a restriction-free summation which also preserves the sum rule. Gritsenko et al.²³ have proposed several corrections (BBCn functional) in order to improve the performance of the BB expression in the calculation of the potential energy curve. All these corrections integrate to zero and therefore the sum rules are satisfied. Unfortunately, the antisymmetry requirements remain violated.

4. The last formula recently published by Piris²⁴ is close to the GU approximation:

$$\begin{aligned} \pi(\mathbf{r}_1, \mathbf{r}_2) = & \sum_i \sum_{j \neq i} n_i n_j \varphi_i^*(\mathbf{r}_1) \varphi_j^*(\mathbf{r}_2) \varphi_i(\mathbf{r}_1) \varphi_j(\mathbf{r}_2) \\ & - \sqrt{n_i n_j} \varphi_i^*(\mathbf{r}_1) \varphi_j^*(\mathbf{r}_2) \varphi_j(\mathbf{r}_1) \varphi_i(\mathbf{r}_2) \quad (20) \\ & - \sum_i^{nco} \sum_{j \neq i}^{nco} \sqrt{(1-n_i)(1-n_j)} \varphi_i^*(\mathbf{r}_1) \varphi_j^*(\mathbf{r}_2) \varphi_j(\mathbf{r}_1) \varphi_i(\mathbf{r}_2) \end{aligned}$$

In eq 20, *nco* denotes the number of HF occupied orbitals. Indeed this formula satisfies none of the requirements.

On the one hand, the evaluation of the ELF function requires the 2-PD and its gradient to be identically zero for $\mathbf{r}_1 = \mathbf{r}_2$, conditions only fulfilled by the HF-XC. In order to calculate the basin pair populations, the approximate formula must obey the sum rules and therefore only the BB or BBCn expression can be retained. There is no approximation suitable for all of the steps in the calculation of the ELF and basin properties. Therefore, in practice, one has to use these two different approximations of the 2-PD: HF-XC is used for the calculation of the ELF values (which yield the ELF basins), whereas BB is used for the calculation of pair populations within the ELF basins (which provide the covariance matrix). It is worth mentioning that a natural-orbital formulation of the ELF equivalent to the first approximation presented here (HF-XC) was suggested in the past by Savin and co-workers (see footnote in ref 30) and more recently by Kohout.⁵

These approximations have been used in the past to approximate the 2-PD in the calculation of the electron sharing indices (ESI) integrated in atomic basins.³¹ In Figure

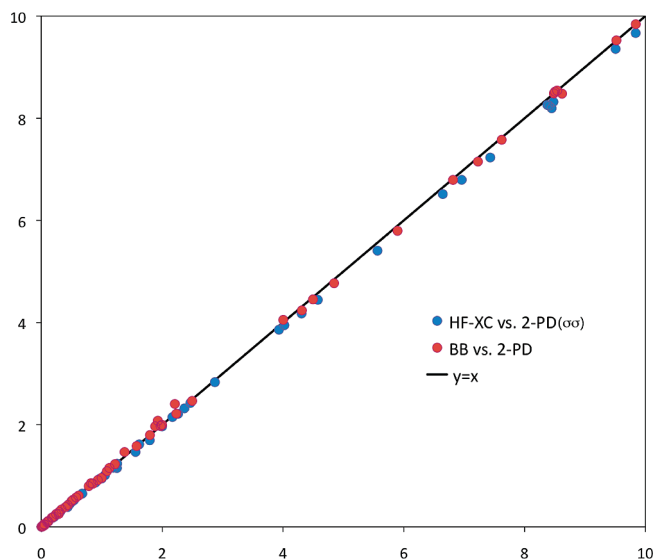


Figure 1. Comparison of ESI (red) and ESI^{σσ} (blue) against the BB and HF-XC approximations for the calculation of 2-PD. The atomic basins correspond to Bader's QTAIM ones.³² From data in ref 31.

1, we show the differences between the *exact*-ESI and those approximated with BB, and the same-spin ESI (which is calculated from the same-spin 2-PD like the ELF) is compared against HF-XC approximation. Interestingly, both approximations are very accurate, in particular, the performance of the HF-like approximation to reproduce the same-spin 2-PD is very satisfactory.

In addition, the aforementioned approximations for the 2-PD can be used for those cases where one cannot obtain the 2-PD. This is the case of, for example, MP2 or CCSD, for which the pair density cannot be calculated as an expectation value, eq 3. In general, variational methods do not suffer this problem, however, it is worth mentioning the case of a truncated configuration interaction (CI) expansion (such as CID or CISD),³³ for which Hellman–Feynman theorem is not fulfilled and thus there are two density representations, the relaxed (obtained as an energy derivative) and the unrelaxed one (obtained as an expectation value), the latter being the only *N*-representable and the one used in the present work.

3. Computational Details

In this work, we have performed HF, DFT, MP2, CCSD, CISD, and CASSCF calculations. Namely, the DFT calculations in this work are performed with Kohn–Sham formalism, and building up an approximate HF-like pair density. For all correlations calculated, we have used HF-XC and BB approximations and the exact 2-PD when possible (CISD, CASSCF).

All calculations were performed with the Gaussian03 program³⁴ using Pople's 6-311+G(2df,2p) basis set with Cartesian d and f orbitals. From the same program the coefficients of the expansion for both CISD and CASSCF were obtained and used by an own program³⁵ to construct the Γ matrix in eq 10. Only the expansion coefficients with values above 10^{-8} were taken, and contributions to the values

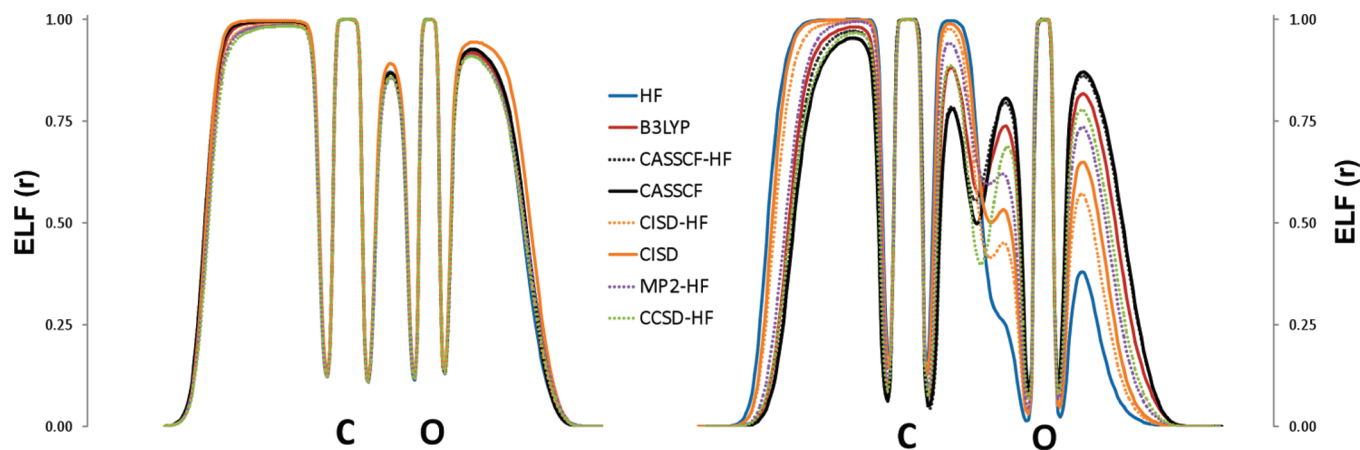


Figure 2. Profile of the electron localization function along the internuclear axis of the carbon monoxide molecule (left: $R = 1.123 \text{ \AA}$; right: $R = 2.0 \text{ \AA}$) for the different methods studied.

of density and pair density above 10^{-12} a.u. have been considered. The wave function together with the Γ matrix were read by our own modification of the ToPMoD package of programs³⁶ to calculate the ELF with both BB and the exact 2-PD. The active space in the CASSCF calculation was composed of the bonding and antibonding valence orbitals and the lone pairs where appropriate. The only exception to this rule is F_2 , for which the experimental bond length³⁷ (1.412 Å) could not be reproduced with the present active space³⁸ (1.460 Å) and it was increased with four orbitals with π -symmetry and prominent 3p character (two bonding and the antibonding counterparts) to obtain a satisfactory agreement with experimental bond length (1.417 Å). The number of grid points in the ELF calculations depends on the size of the molecule. A 3D box containing the molecule within 5 Å to the limits of the box is constructed. The number of points are distributed in each of the edges in such a way that they are separated 0.1 Å from each other. For example, in the case of water molecule at the B3LYP level, the number of grid points is $100 \times 128 \times 110$.

4. Numerical Validation

Figure 2 illustrates how the HF-XC approximation works for the different methods. The calculations have been carried out on the carbon monoxide molecule for two internuclear separations.

At the equilibrium distance, all functions yield results in very good agreement. At a larger C–O distance, 2 Å, the effects of the electron correlation are more important: at all levels, the curve presents a secondary maximum between C and O; the only exception is HF, where we can appreciate a shoulder. CISD, CISD-HF (i.e., CISD with the 2-PD constructed using the CISD NOs, see eq 16), and MP2-HF improve HF, but exhibit only a small minimum. This can be attributed to the small amount of electron correlation introduced by these methods. CCSD-HF and B3LYP qualitatively reproduce the topological features of the CASSCF localization function. However, they tend to underestimate the function value at the secondary maximum in the C–O bonding region, as well as at the attractor of the V(O) basin

Table 1. Populations and Same Spin Pair Populations of the CO Localization Basins Calculated with the Exact CASSCF Pair Function (Exact) and with Buijse and Baerends (BB) Approximation

Ω	\bar{N}_Ω		$\bar{N}_{\Omega\Omega}^{\uparrow\uparrow}$	
	exact	BB	exact	BB
C(C)	2.07	2.07	0.15	0.15
C(O)	2.10	2.10	0.22	0.22
V(C,O)	3.13	3.10	1.60	1.53
V(C)	2.50	2.51	0.68	0.66
V(O)	4.19	4.22	2.99	2.98

on the right side. Interestingly, CASSCF-HF curve mimics the CASSCF one.

Table 1 shows that the pair populations calculated with the Buijse and Baerends pair function are very close to the exact CASSCF pair function ones. Surprisingly, there are some small deviations for the populations themselves which might be due to numerical round off in both the natural orbitals and the exact pair function determinations.

5. Results

In order to assess how the HF-XC and BB approximation work in the ELF, we have chosen a set of molecules, NH_3 , CO_2 , H_2O , CO , CN^- , NO^+ , N_2 , H_2O_2 , and F_2 , which have been calculated using a plethora of methods: HF, B3LYP, MP2, CISD, CCSD, and CASSCF. The data has been divided in three groups of molecules: those with small electron-correlation effects, Table 2, species with moderate electron-correlation influence, Table 3, and molecules showing larger electron-correlation effects, Table 4.

First of all, let us analyze the basin populations. Basin populations are calculated through integration of the electron density into the ELF basins. Thus, these numbers are only affected by HF-XC approximation in the calculation of the ELF basins. In those cases where electron correlation affects the electron population, such as CO_2 or the molecules in Tables 3 and 4, the HF-XC approach gives values close to the *exact* expression. The larger differences in the populations with respect to CASSCF values are found for CISD, MP2-BB, and CISD-BB. Interestingly, in those cases where CISD values differ most from the CASSCF ones, HF-XC usually

Table 2. Set of Molecules Showing Small Electron-Correlation Effects: NH₃, CO₂, and H₂O

		HF	B3LYP	MP2-BB	CISD-BB	CISD	CCSD-BB	CASSCF-BB	CASSCF
NH ₃	r_{NH}	0.998	1.013	1.011	1.008	1.008	1.011	1.022	1.022
	$\angle HNH$	108.2	107.5	107.1	107.1	107.1	106.9	105.1	105.1
	V(N)	2.10	2.16	2.14	2.15	2.18	2.15	2.21	2.19
	$\sigma^2(N)$	0.92	0.95	0.87	0.87	0.87	0.86	0.94	0.94
	V(N,H)	1.93	1.91	1.91	1.91	1.90	1.91	1.89	1.90
	$\sigma^2(N,H)$	0.75	0.76	0.68	0.68	0.72	0.67	0.71	0.74
H ₂ O	r_{OH}	0.940	0.961	0.959	0.954	0.954	0.957	0.962	0.962
	$\angle HOH$	106.4	105.2	104.2	104.8	104.8	104.5	104.7	104.7
	V(O)	2.23	2.28	2.27	2.27	2.28	2.28	2.26	2.25
	$\sigma^2(O)$	1.04	1.06	0.98	0.98	0.99	0.97	1.03	1.01
	V(H,O)	1.71	1.66	1.68	1.68	1.67	1.67	1.69	1.69
	$\sigma^2(H,O)$	0.79	0.79	0.71	0.71	0.75	0.70	0.75	0.78
CO ₂	r_{CO}	1.136	1.161	1.170	1.152	1.152	1.161	1.164	1.164
	V(O)	5.71	5.18	5.19	5.21	5.22	5.22	5.25	5.20
	$\sigma^2(O)$	1.39	1.42	1.19	1.22	1.38	1.18	1.25	1.36
	V(C,O)	2.15	2.67	2.67	2.64	2.63	2.63	2.60	2.66
	$\sigma^2(C,O)$	1.19	1.38	1.28	1.28	1.47	1.26	1.32	1.38

Table 3. Set of Isoelectronic Molecules with Moderate Electron-Correlation Effects: CO, CN⁻, NO⁺, and N₂

		HF	B3LYP	MP2-BB	CISD-BB	CISD	CCSD-BB	CASSCF-BB	CASSCF
CO	r_{CO}	1.103	1.125	1.136	1.121	1.121	1.127	1.132	1.132
	V(C)	2.37	2.54	2.58	2.47	2.49	2.50	2.51	2.50
	$\sigma^2(C)$	0.66	0.77	0.65	0.59	0.70	0.56	0.67	0.71
	V(C,O)	3.28	3.04	3.08	3.15	2.96	3.10	3.10	3.13
	$\sigma^2(C,O)$	1.44	1.40	1.29	1.31	1.33	1.29	1.36	1.41
	V(O)	4.16	4.24	4.16	4.20	4.37	4.22	4.22	4.19
	$\sigma^2(O)$	1.39	1.41	1.17	1.19	1.29	1.16	1.27	1.33
CN ⁻	r_{CN}	1.152	1.171	1.189	1.169	1.169	1.176	1.186	1.186
	V(N)	3.30	3.49	3.50	3.42	3.52	3.44	3.47	3.44
	$\sigma^2(N)$	1.20	1.26	1.02	1.04	1.16	1.00	1.10	1.16
	V(N,C)	3.86	3.46	3.41	3.62	3.46	3.56	3.51	3.54
	$\sigma^2(N,C)$	1.52	1.47	1.32	1.36	1.39	1.33	1.43	1.48
	V(C)	2.64	2.83	2.91	2.76	2.82	2.80	2.83	2.83
	$\sigma^2(C)$	0.87	0.98	0.79	0.76	0.89	0.73	0.84	0.89
NO ⁺	r_{NO}	1.026	1.057	1.083	1.051	1.051	1.059	1.065	1.065
	V(N)	2.65	2.85	2.98	2.79	2.83	2.83	2.86	2.85
	$\sigma^2(N)$	0.87	0.91	0.81	0.77	0.90	0.75	0.83	0.89
	V(N,O)	3.57	3.15	2.98	3.25	3.09	3.18	3.14	3.17
	$\sigma^2(N,O)$	1.49	1.42	1.27	1.33	1.34	1.31	1.37	1.42
	V(O)	3.62	3.82	3.85	3.77	3.91	3.79	3.81	3.79
	$\sigma^2(O)$	1.29	1.35	1.08	1.11	1.24	1.08	1.15	1.22
N ₂	r_{NN}	1.066	1.091	1.113	1.088	1.088	1.096	1.103	1.103
	V(N)	2.96	3.17	3.24	3.11	3.18	3.14	3.17	3.15
	$\sigma^2(N)$	1.05	1.13	0.92	0.92	1.04	0.90	0.97	1.02
	V(N,N)	3.89	3.49	3.31	3.60	3.48	3.52	3.48	3.52
	$\sigma^2(N,N)$	1.53	1.48	1.31	1.36	1.42	1.34	1.42	1.48

gives values closer to those given by CASSCF; see, for instance, CO or V(O) basin in NO⁺. Although this finding is to be regarded to some extent as fortuitous, the significance of this fact must not be overlooked. As a particularly difficult molecule, one may mention F₂ for which B3LYP or CISD attribute a large population to the bonding basin. It is noteworthy that CCSD and CASSCF-BB populations are quite close to the exact CASSCF values. In general, the HF-XC performs very satisfactorily, and the differences found in the populations are mostly attributed to the amount of electron-correlation introduced by the method rather than to the approximation used for the calculation of the 2-PD. It is noteworthy that in the performance of B3LYP; in all cases, with the exception of F₂, the populations obtained are very close to CASSCF. The ELF topology of N₂ and F₂ was recently reviewed by Bezugly et al.³⁹ using HF and MRCI

methods. However, no population analysis was performed in this study.

In the following, we will analyze the performance of BB-approximation by comparing the values of the variance of the electron population within the ELF basin, σ^2 . One should take into account that these values are also affected by the HF-XC approximation, used to calculate the basin boundaries, and therefore one anticipates further differences than those found in the case of the electron populations.

The values of σ^2 are sensibly smaller for the methods that use an approximate 2-PD in the calculation of the variance, even for those molecules not much affected by electron correlation. The fact that the variance is lower with the BB-approximation was already observed in QAIM partition of the molecular space. See ref 31, where $\lambda(A)$ was shown to be usually larger with the BB-approximation. Since $\sigma^2(A)$

Table 4. Molecules with Large Electron-Correlation Effects: H₂O₂ and F₂^a

		HF	B3LYP	MP2-BB	CISD-BB	CISD	CCSD-BB	CASSCF-BB	CASSCF
H ₂ O ₂	r_{OO}	1.385	1.448	1.449	1.417	1.417	1.439	1.467	1.467
	r_{OH}	0.942	0.966	0.965	0.954	0.954	0.961	0.968	0.968
	\angle HOO	103.2	100.8	99.8	101.5	101.5	100.7	99.8	99.8
	\angle HOOH	112.0	114.9	115.4	113.3	113.3	114.1	114.0	114.0
	$V(\text{H},\text{O})$	1.75	1.69	1.69	1.71	1.70	1.70	1.68	1.70
	$\sigma^2(\text{H},\text{O})$	0.78	0.78	0.71	0.71	0.75	0.70	0.72	0.77
	$V(\text{O},\text{O})$	0.81	2 × 0.30	2 × 0.28	0.67	0.67	2 × 0.30	2 × 0.28	0.60
	$\sigma^2(\text{O},\text{O})$	0.60	0.53	0.47	0.50	0.51	0.51	0.50	0.48
	$V(\text{O})$	2.37	2.45	2.46	2.42	2.43	2.45	2.46	2.44
F ₂	$\sigma^2(\text{O})$	1.04	1.08	0.98	0.98	1.03	0.97	1.04	1.05
	r_{FF}	1.326	1.395	1.396	1.368	1.368	1.392	1.417	1.417
	$V(\text{F})$	6.53	6.63	6.66	6.60	6.62	6.62	6.66	6.72
	$\sigma^2(\text{F})$	0.99	0.98	0.76	0.81	0.92	0.75	0.77	0.86
	$V(\text{F},\text{F})$	0.57	2 × 0.25	2 × 0.17	2 × 0.19	2 × 0.23	2 × 0.18	2 × 0.19	2 × 0.15
	$\sigma^2(\text{F},\text{F})$	0.46	0.45	0.28	0.32	0.41	0.30	0.31	0.29

^a If a given method splits the bonding basin into two basins, then for the sake of comparison, the population variance is given for the basin resulting of merging the two bonding basins.

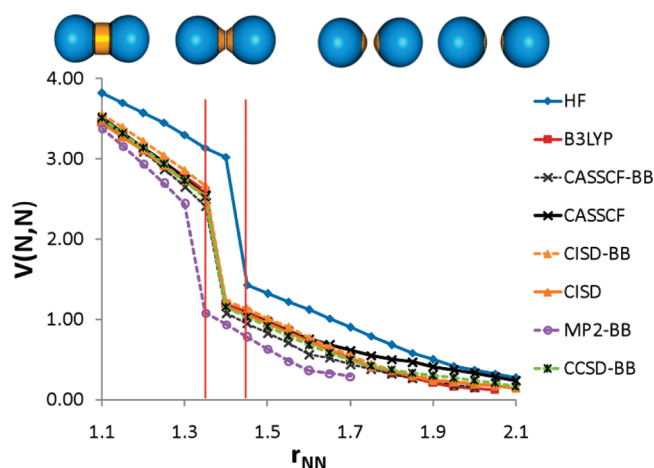


Figure 3. The population of the bonding basin of N₂ along the dissociation curve. When the basin splits into two, only the population of one of them is represented.

$= N(A) - \lambda(A)$, and populations are almost invariant, one expects BB-approximated σ^2 values to be smaller than the exact ones. If we compare the values of CISD, CISD-BB, CCSD-BB, and CASSCF-BB against the values of CASSCF, then we conclude that CISD performs best closely followed by CASSCF-BB, CISD-BB, and CCSD-BB. Although CISD-BB and CCSD-BB perform similarly, it is surprising that CISD-BB is systematically better than CCSD-BB. It indicates that BB-approximation is not particular fortunate in the case of CCSD. Not unexpectedly, MP2 overcorrects HF results, in line with the well-known fact that MP2 overestimates correlation.^{40,41} On the other hand, B3LYP results are very good, performing close to CASSCF-BB, with only some exceptions for those systems where the influence of electron-correlation is more notorious, and the adequacy of B3LYP functional for the correct description of both the geometry and the electronic structure is questionable.

Finally, we have calculated the ELF along the dissociation curve of N₂. Figures 3 and 4 show the change in the bonding basins population and variance, respectively, along the reaction path. Upon dissociation, the bonding basin is split into two basins, which are later absorbed by the lone-pair basin to form a spherical basin corresponding to the valence

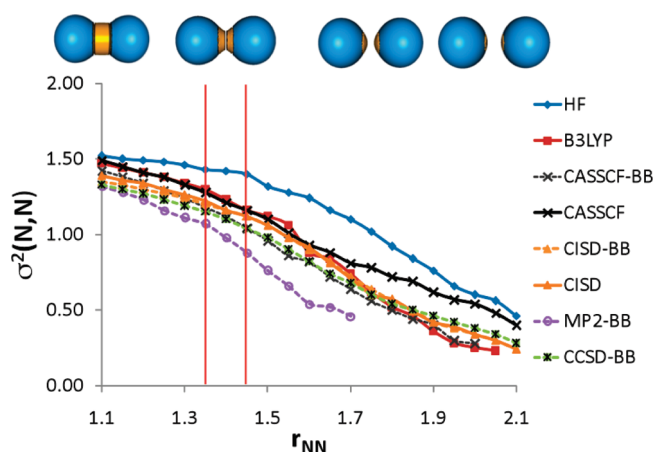


Figure 4. The variance of the population of the bonding basin of N₂ along the dissociation curve. As the molecule stretches the lone-pair basin absorbs the bonding basin.

shell of N atom. In general, all methods give a similar population along the reaction path with the exception of HF and MP2 which split the bonding basin too late and too early, respectively. In addition, at the MP2 level the bonding basin is absorbed by the lone-pair much sooner (at 1.7 Å). By analyzing the variance of this bonding basin, we reach a similar conclusion. Only after 1.7 Å appreciable differences between CASSCF and the other methods are encountered. On this point and further, CCSD-BB and CISD perform slightly better than the rest.

6. Conclusions

In this work we have presented a 2-fold approximation for the calculation of the ELF which avoids the use of the two-particle density (2-PD). The approximation only needs the natural orbitals and their occupancies, which are available for most methods used in electronic structure calculations. In this way, methods such as CCSD and MP2 can be used for the calculation of the ELF despite the lack of a pertinent definition of the 2-PD. The first approximation (HF-XC) relies on the single-determinant approximation for the calculation of the 2-PD and it is used for the calculation of the ELF itself (such approximation was suggested in the past^{5,30}). The performance of this approximation is extremely

good, as shown in a number of examples. The second approximation is used for the calculation of pair densities integrated in the ELF basins. It also relies on both natural orbitals and occupancies, but uses a different expression due to Müller and popularized by Baerends and Buijse. The performance of this approximation is also very convincing for CASSCF calculations but systematically underestimates the values of the variance for CCSD calculations. B3LYP exhibits a performance close enough to the approximate CASSCF, except for those molecules where B3LYP is an inappropriate choice for the description of system. The present formulation avoids the calculation of the 2-PD, thus providing the means for routine calculations of the ELF for medium-size molecules with correlated methods.

Acknowledgment. F.F. thanks the MICINN for a doctoral fellowship, No. AP2005-2997. E.M. acknowledges financial support from Marie Curie IntraEuropean Fellowship, Seventh Framework Programme (FP7/2007–2013), under Grant Agreement No. PIEF-GA-2008-221734 and from the Polish Ministry of Science and Higher Education (Project No. N N204 215634). M.S. and F.F. are grateful for financial help furnished by the Spanish MICINN Project No. CTQ2008-03077/BQU and by the Catalan DIUE through Project No. 2009SGR637. Support for the research of M. S. was received through the ICREA Academia prize for excellence in research funded by the DIUE of the Generalitat de Catalunya. We thank the Centre de Supercomputació de Catalunya (CESCA) for partial funding of computer time.

References

- Matito, E.; Silvi, B.; Duran, M.; Solà, M. *J. Chem. Phys.* **2006**, *125*, 024301.
- Becke, A. D.; Edgecombe, K. E. *J. Chem. Phys.* **1990**, *92*, 5397–5403.
- Silvi, B. *J. Phys. Chem. A* **2003**, *107*, 3081–3085.
- Kohout, M. *Int. J. Quantum Chem.* **2004**, *97*, 651–658.
- Kohout, M.; Pernal, K.; Wagner, F. R.; Grin, Y. *Theor. Chem. Acc.* **2004**, *112*, 453–459.
- Kohout, M.; Pernal, K.; Wagner, F. R.; Grin, Y. *Theor. Chem. Acc.* **2005**, *113*, 287–293.
- Wagner, F. R.; Bezugly, V.; Kohout, M.; Grin, Y. *Chem.—Eur. J.* **2007**, *13*, 5724–5741.
- Noury, S.; Krokidis, X.; Fuster, F.; Silvi, B. *Comput. Chem.* **1999**, *23*, 597–604.
- Matito, E.; Solà, M. *Coord. Chem. Rev.* **2009**, *253*, 647–665.
- Poater, J.; Duran, M.; Solà, M.; Silvi, B. *Chem. Rev.* **2005**, *105*, 3911–3947.
- Santos, J. C.; Tiznado, W.; Contreras, R.; Fuentealba, P. *J. Chem. Phys.* **2004**, *120*, 1670–1673.
- Santos, J. C.; Polo, V.; Andrés, J. *Chem. Phys. Lett.* **2005**, *406*, 393–397.
- Krokidis, X.; Noury, S.; Silvi, B. *J. Phys. Chem. A* **1997**, *101*, 7277–7282.
- Berski, S.; Andrés, J.; Silvi, B.; Domingo, L. *J. Phys. Chem. A* **2003**, *107*, 6014–6024.
- Matito, E.; Solà, M.; Duran, M.; Poater, J. *J. Phys. Chem. B* **2005**, *109*, 7591–7593.
- Matito, E.; Poater, J.; Duran, M.; Solà, M. *Chem. Phys. Chem.* **2006**, *7*, 111–113.
- Smith, D. W. *J. Chem. Phys.* **1965**, *43*, S258–S264.
- Barnett, G. P.; Shull, H. *Phys. Rev.* **1967**, *153*, 61–73.
- Müller, A. M. K. *Phys. Lett.* **1984**, *105A*, 446–452.
- Goedecker, S.; Umrigar, C. J. *Phys. Rev. Lett.* **1998**, *81*, 866–869.
- Kutzelnigg, W.; Mukherjee, D. *J. Chem. Phys.* **1999**, *110*, 2800–2809.
- Buijse, M. A.; Baerends, E. J. *Mol. Phys.* **2002**, *100*, 401–421.
- Gritsenko, O.; Pernal, K.; Baerends, E. J. *J. Chem. Phys.* **2005**, *122*, 204102.
- Piris, M. *Int. J. Quantum Chem.* **2006**, *106*, 1093–1104.
- Buijse, M. A. Ph.D. thesis, Vrije Universiteit: Amsterdam, The Netherlands, 1991.
- McWeeny, R. *Methods of Molecular Quantum Mechanics*, 2nd ed.; Academic Press: London, 1989.
- Kato, T. *Comm. Pure Appl. Math* **1957**, *10*, 151–177.
- Rajagopal, A. K.; Kimball, J. C.; Banerjee, M. *Phys. Rev. B* **1978**, *18*, 2339–2345.
- Holas, A. *Phys. Rev. A* **1999**, *59*, 3454–3461.
- Savin, A.; Silvi, B.; Colonna, F. *Can. J. Chem.* **1996**, *74*, 1088–1096.
- Matito, E.; Solà, M.; Salvador, P.; Duran, M. *Faraday Discuss.* **2007**, *135*, 325–345.
- Bader, R. F. W. *Atoms in Molecules: A Quantum Theory*; Oxford University Press: Oxford, 1990.
- Raghavachari, K.; Pople, J. A. *Int. J. Quantum Chem.* **1981**, *20*, 1067–1071.
- Frisch, M. J. et al. *Gaussian 03, Revision C.02*, Gaussian, Inc., Pittsburgh, PA, 2003.
- Matito, E.; Feixas, F. *DMn program*, 2009, University of Girona (Spain) and University of Szczecin (Poland).
- Noury, S.; Krokidis, X.; Fuster, F.; Silvi, B. *ToPMoD package* **1997**, Université Pierre et Marie Curie (France).
- Herzberg, G.; Huber, K. P. *Molecular Spectra and Molecular Structure. IV. Constants of Diatomic Molecules*; Van Nostrand: New York, 1979.
- Peterson, K. A.; Kendall, R. A.; Dunning, T. H. *J. Chem. Phys.* **1993**, *99*, 9790–9805.
- Bezugly, V.; Wielgus, P.; Kohout, M.; Wagner, F. R. *J. Comput. Chem.* **2010**, *31*, 2273–2285.
- Torrent, M.; Gili, P.; Duran, M.; Solà, M. *J. Chem. Phys.* **1996**, *104*, 9499–9510.
- Wang, J.; Shi, Z.; Boyd, R. J.; Gonzalez, C. A. *J. Phys. Chem.* **1994**, *98*, 6988–6994.

On the Dissociation of Ground State *trans*-HOOO Radical: A Theoretical Study

Josep M. Anglada,[†] Santiago Olivella,^{*,†} and Albert Solé[‡]

Institut de Química Avançada de Catalunya, CSIC, Jordi Girona 18-26, 08034-Barcelona, Catalonia, Spain, and Departament de Química Física and Institut de Química Teòrica i Computacional, Universitat de Barcelona, Martí i Franquès 1, 08028-Barcelona, Catalonia, Spain

Received June 28, 2010

Abstract: The hydrotrioxyl radical (HOOO[•]) plays a crucial role in atmospheric processes involving the hydroxyl radical (HO[•]) and molecular oxygen (O₂). The equilibrium geometry of the electronic ground state (X ²A'') of the *trans* conformer of HOOO[•] and its unimolecular dissociation into HO[•] (X ²Π) and O₂ (X ³Σ_g⁻) have been studied theoretically using CASSCF and CASPT2 methodologies with the aug-cc-pVTZ basis set. On the one hand, CASSCF(19,15) calculations predict for *trans*-HOOO[•] (X ²A'') an equilibrium structure showing a central O–O bond length of 1.674 Å and give a classical dissociation energy $D_e = 1.1$ kcal/mol. At this level of theory, it is found that the dissociation proceeds through a transition structure involving a low energy barrier of 1.5 kcal/mol. On the other hand, CASPT2(19,15) calculations predict for *trans*-HOOO[•] (X ²A'') a central O–O bond length of 1.682 Å, which is in excellent agreement with the experimental value of 1.688 Å, and give $D_e = 5.8$ kcal/mol. Inclusion of the zero-point energy correction (determined from CASSCF(19,15)/aug-cc-pVTZ harmonic vibrational frequencies) in this D_e leads to a dissociation energy at 0 K of $D_0 = 3.0$ kcal/mol. This value of D_0 is in excellent agreement with the recent experimentally determined $D_0 = 2.9 \pm 0.1$ kcal/mol of Le Picard et al. (*Science* **2010**, *328*, 1258–1262). At the CASPT2 level of theory, we did not find for the dissociation of *trans*-HOOO[•] (X ²A'') an energetic barrier other than that imposed by the endoergicity of the reaction. This prediction is in accordance with the experimental findings of Le Picard et al., indicating that the reaction of HO[•] with O₂ yielding HOOO[•] is a barrierless association process.

1. Introduction

The hydrotrioxyl radical (HOOO[•]) is an important reactive intermediate of interest in many areas of chemistry, mainly in atmospheric chemistry.^{1–23} Its equilibrium geometry, vibrational frequencies, reactivity, and the binding energy of the central O–O bond (designated by HO–OO[•]) have been the subject of numerous experimental and theoretical studies,^{6,8–11,19,24–48} which have been recently summarized by Lester and co-workers.⁴⁹ The most important effect of HOOO[•] in the Earth's atmosphere would arise if it could

serve as a temporary sink for hydroxyl radicals (HO[•]) after their association with molecular oxygen (O₂).⁵⁰

The first direct experimental observation of HOOO[•] came from the work by Cacace et al.³² using neutralization–reionization mass spectrometry, and it was subsequently observed by infrared spectroscopy in Ar and H₂O-ice matrices.³⁴ Spectroscopic observations of rotational transitions of HOOO[•] and DOOO[•] in the gas phase were made by Suma et al.²⁴ using Fourier-transform microwave (FTMW) spectroscopy. The determined rotational constants, as well as the ratio of *a*-type and *b*-type transition intensities, were consistent with a *trans* planar structure of ²A'' character; however, the *cis* conformer was not observed. In conjunction with high-level multireference configuration interaction

* Corresponding author e-mail: sonqtc@cid.csic.es.

[†] Institut de Química Avançada de Catalunya.

[‡] Universitat de Barcelona.

(MRCI) calculations, the *A* and *B* rotational constants were used to determine an equilibrium geometry in which the HO–OO[•] bond length was 1.688 Å. Concerning the binding energy of this markedly long O–O bond, there has been much debate in the literature as to whether ground state (X^2A'') of *trans*-HOOO[•] is sufficiently stable relative to the HO[•] ($X^2\Pi$) + O₂ ($X^3\Sigma_g^-$) dissociation limit to be present in measurable concentrations in the Earth's atmosphere. A full assessment of the atmospheric abundance of HOOO[•] requires detailed knowledge of its thermochemical properties. An indirect measurement of the stability of HOOO[•] was made by Speranza,³¹ who inferred an enthalpy of formation of HOOO[•] of -1 ± 5 kcal/mol and consequently a HO–OO[•] bond dissociation enthalpy of 10 ± 5 kcal/mol. This value has since been re-examined by experiments reported by Lester and co-workers^{1,2} using IR–UV double resonance spectroscopy to measure the HO[•] product state distribution following vibrational predissociation of HOOO[•]. From the energetically highest observed HO[•] product channel, an upper limit dissociation energy at 0 K (designated by D_0) of 6.12 kcal/mol was established for the HO–OO[•] bond.¹ An analogous study of the DOOO[•] isotopomer allowed this value to be refined and reduced the upper limit to 5.31 kcal/mol.⁵⁰ In addition, the latter value could be further improved if accurate zero-point vibrational energies for HOOO[•] and DOOO[•] were known.⁵¹

In comparing the experimentally determined upper limit of 5.31 kcal/mol for D_0 with the values obtained from theoretical calculations, the agreement between experimental and theoretical thermochemistry for HOOO[•] is poor. As noted by Varner et al.,⁴⁷ recent calculations of the D_0 in the *trans*-HOOO[•] conformer have begun to cluster around 1 to 3 kcal/mol, significantly below the experimental upper bound of 5.31 kcal/mol. One exception is the DFT calculations reported by Braams and Yu⁴² based on the HCTH functional in conjunction with the aug-cc-pVTZ basis set, which predict $D_0 = 6.15$ kcal/mol. However, it should be pointed out that these HCTH/aug-cc-pVTZ calculations predict a HO–OO[•] bond length of 1.610 Å for *trans*-HOOO[•], significantly shorter than the value of 1.688 Å derived from the experimental rotational constants.²⁴

Varner et al.⁴⁷ have suggested that the apparent discrepancy between theory and experiment on the magnitude of the HO–OO[•] binding energy in the *trans*-HOOO[•] may be rationalized in terms of the existence of an exit barrier for the unimolecular dissociation. Using an equation-of-motion coupled-cluster singles and doubles method (designated by EOMIP-CCSD*) with the cc-pVQZ basis set, Varner et al.⁴⁷ investigated the dissociation profile of *trans*-HOOO[•] by computing the minimum energy for fixed values of the HO–OO[•] bond. In agreement with earlier theoretical studies^{8,19,26,27,35} of the HOOO[•] dissociation to HO[•] + O₂, an exit barrier of approximately 3–5 kcal/mol was found. On the basis of when a barrier exists in a unimolecular decomposition pathway, highly nonstatistical behavior can occur; Varner et al.⁴⁷ suggest that it would seem plausible that perhaps 2 kcal/mol could be carried off by translational degrees of freedom. This would tend to greatly offset the difference between the theoretical D_0 value of 2–3 kcal/

mol and the experimental upper limit of 5.31 kcal/mol determined by Lester and co-workers.⁵⁰

During the preparation of this paper, Le Picard et al.⁵² reported an experimental study on the decay of HO[•] radicals in the presence of O₂ at low temperatures (55.7–110.8 K) in a supersonic flow apparatus. Their study enabled the derivation of an experimental D_0 of 2.9 ± 0.1 kcal/mol, significantly below the experimental upper bound of 5.31 kcal/mol determined by Lester and co-workers.⁵⁰ In addition, the third-order rate constants for HOOO[•] formation determined in the study of Le Picard et al.⁵² were found to have a strong negative temperature dependence, indicative of a barrierless association reaction. Therefore, there is a clear discrepancy between the experimental findings of Le Picard et al.⁵² and the theoretical prediction of Varner et al.⁴⁷ concerning the existence of an exit barrier for the unimolecular dissociation of *trans*-HOOO[•] (X^2A'') to HO[•] ($X^2\Pi$) + O₂ ($X^3\Sigma_g^-$).

Since it has been argued^{10,11,24,41,42,44} that multireference-based methods are needed to correctly describe the electronic structure of HOOO[•], the results of Varner et al.⁴⁷ motivated us to reinvestigate the minimum energy reaction path of *trans*-HOOO[•] (X^2A'') dissociation to HO[•] ($X^2\Pi$) + O₂ ($X^3\Sigma_g^-$) using high level multireference-based methodologies. With this aim, herein, we report the results of CASSCF and CASPT2 electronic structure calculations on the equilibrium geometry of ground-state *trans*-HOOO[•] and its unimolecular dissociation.

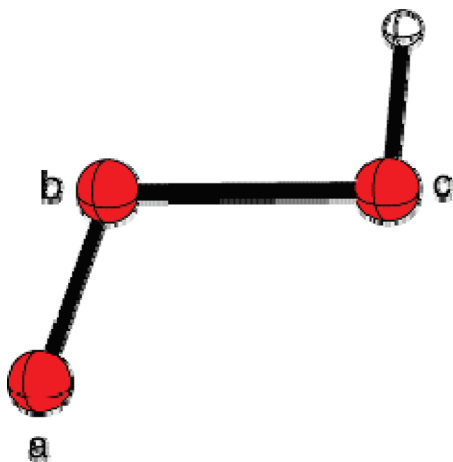
2. Results and Discussion

2.1. CASSCF Approach. The calculations reported in this section were performed within the framework of the multiconfigurational self-consistent field (MCSCF) methodology of the complete active space (CAS) SCF class⁵³ employing Dunning's augmented correlation-consistent polarized valence triple- ζ (aug-cc-pVTZ) basis set.⁵⁴ For *trans*-HOOO[•], all 19 valence electrons were distributed in all possible ways in the active space generated following the procedure suggested by Anglada and Bofill,⁵⁵ based on the fractional occupation of the natural orbitals generated from the first-order density matrix calculated from an initial multireference single- and double-excitation configuration interaction (MRDCI) wave function correlating all valence electrons. These indicated a CAS of 15 orbitals (11 a' and 4 a'') and led to a CASSCF wave function, designated by CASSCF(19,15), formed as a linear combination of 6 246 240 doublet spin-adapted configuration state functions (CSFs). This CASSCF wave function has already been used with the 6-311+G(2df,2p) basis set in a recent⁴¹ theoretical study on the nature of the long O–O bond in HOOO[•]. For the dissociation products of *trans*-HOOO[•] (X^2A''), the aforementioned active space selection procedure indicated a number of active orbitals that varied from seven, for HO[•] ($X^2\Pi$), to eight, for O₂ ($X^3\Sigma_g^-$). The distribution among these active orbitals of the corresponding number of valence electrons led to the CASSCF wave functions denoted by CASSCF(7,7) and CASSCF(12,8), respectively.

Table 1. Relative Energies^a (kcal/mol) and Geometrical Parameters^b (distances in Å and angles in deg) of the Relevant Stationary Points on the Ground-State Potential Energy Surface for the Dissociation of *trans*-HOOO* (X²A'') to HO* (X²Π) + O₂ (X³Σ_g⁻) Calculated at the CASSCF(19,15)/aug-cc-pVTZ Level

species	ΔU	ΔE_0	$r(\text{O}_a\text{O}_b)$	$r(\text{O}_b\text{O}_c)$	$r(\text{O}_c\text{H})$	$\theta(\text{O}_a\text{O}_b\text{O}_c)$	$\theta(\text{O}_b\text{O}_c\text{H})$	$\tau(\text{O}_a\text{O}_b\text{O}_c\text{H})$
<i>trans</i> -HOOO*	0.0	0.0	1.226	1.674	0.973	109.9	96.9	180.0
TS	1.5	0.1	1.214	2.128	0.974	112.0	90.5	176.4
CX	0.9	-1.4	1.217	2.894	0.974	114.0	84.9	179.6
HO* + O ₂	1.1	-1.7	1.218		0.974			

^a ΔU is the relative electronic energy, and ΔE_0 is the ZPVE-corrected relative energy. ^b Geometrical parameters defined as in Figure 1.

**Figure 1.** Optimized equilibrium structure of *trans*-HOOO* (X²A'').

The geometries of the relevant stationary points (minima and first-order saddle points) on the lowest-energy potential energy surface (PES) of the dissociation of *trans*-HOOO* (X²A'') to HO* (X²Π) + O₂ (X³Σ_g⁻) were optimized at the CASSCF(19,15) level of theory by using analytical gradients procedures.⁵⁶ The harmonic vibrational frequencies of these stationary points were computed at the same level of theory. Zero-point vibrational energies (ZPVEs) were determined from unscaled CASSCF(19,15) harmonic vibrational frequencies. All of the CASSCF calculations were carried out using the GAMESS⁵⁷ program package.

In Table 1, we present the relative electronic energies (designated by ΔU), the ZPVE-corrected relative energies (designated by ΔE_0), and geometrical parameters of the stationary points (see Figure 1 for atom labeling). The calculated harmonic vibrational frequencies are compared with the vibrational frequencies measured from the experimental IR spectra in Table S4 (Supporting Information). Overall, the geometrical parameters computed for *trans*-HOOO* compare well with the values determined from the experimental rotational constants in conjunction with MRCI calculations.²⁴ In particular, the HO–OO* bond length ($r(\text{O}_b\text{O}_c)$) of 1.674 Å and the terminal O–O bond length ($r(\text{O}_a\text{O}_b)$) of 1.226 Å are in good agreement with the experimental values of 1.688 and 1.225 Å, respectively.²⁴ However, the $\theta(\text{O}_b\text{O}_c\text{H})$ bond angle is calculated to be 6.9° larger than the experimental value of 90.0°. On the other hand, our CASSCF(19,15) calculations largely underestimate the binding energy of the HO–OO* bond. That is, whereas the ΔU data listed in Table 1 indicate that *trans*-HOOO* lies 1.1 kcal/mol below the energy of HO* (X²Π) + O₂ (X³Σ_g⁻), the calculated ΔE_0 data show that *trans*-HOOO* is

Table 2. A Comparison of Barrier Heights (in kcal/mol) Calculated for the Dissociation of HOOO* into HO* + O₂

method	ΔU^\ddagger ^a	ΔE_0^\ddagger ^b	ΔH^\ddagger (298 K) ^c	ref
CASSCF(3,3)	4			26
MCHF	16.5			27
QCISD(T)-CBS	8.9			35
B3LYP			3.7	19
CCSD(T)-CBS(W1U)	6.6	5.3		8
EOMIP-CCSD*	3–4			47
CASSCF(19,15)	1.5	0.1		this work

^a Potential energy barrier. ^b Energy barrier at 0 K. ^c Enthalpy barrier at 298 K.

1.7 kcal/mol more energetic than the dissociation products. These results are in clear disagreement with the most recent experimental D_0 of 2.9 ± 0.1 kcal/mol.⁵²

Our CASSCF(19,15) calculations predict that the dissociation of *trans*-HOOO* (X²A'') to HO* (X²Π) + O₂ (X³Σ_g⁻) proceeds via a transition structure. The geometrical parameters computed for this transition structure, labeled as **TS**, are shown in Table 1. From the geometry point of view, the main feature of **TS** is the long $r(\text{O}_b\text{O}_c)$ distance of 2.128 Å. The ΔU data listed in Table 1 show that the dissociation of *trans*-HOOO* involves a potential energy barrier (designated by ΔU^\ddagger) of 1.5 kcal/mol. Inclusion of ZPVE corrections to energy leads to an energy barrier at 0 K (designated by ΔE_0^\ddagger) of only 0.1 kcal/mol. These barrier heights are compared in Table 2 with the values found in previous theoretical studies. In particular, it is remarkable that the ΔU^\ddagger of 1.5 kcal/mol computed at the CASSCF(19,15) level is much lower than the ΔU^\ddagger of 6.6 kcal/mol reported by Fabian et al.⁸ based on CCSD(T)-CBS (W1U) calculations and significantly lower than the ΔU^\ddagger of approximately 3–4 kcal/mol estimated by Varner et al.⁴⁷ at the EOMIP-CCSD* level. This is attributed to the fact that both CCSD(T)-CBS (W1U) and EOMIP-CCSD* are correlated single-reference-based methods, which take into account the dynamic electron correlation effects, whereas CASSCF(19,15) is a multiconfigurational method, which takes into account the near degeneracy effects in the electronic structure (i.e., the nondynamic electron correlation).

Intrinsic reaction coordinate calculations⁵⁸ showed that **TS** goes backward to *trans*-HOOO* and goes forward to give a product complex in which the HO* radical is loosely bound to the O₂. The optimized geometry of this complex, labeled as **CX**, was characterized as a true local minimum on the PES. There is no classical barrier for the dissociation of **CX** into HO* (X²Π) + O₂ (X³Σ_g⁻) other than that imposed by the endoergicity of the process. The relative energies and geometrical parameters computed for **CX** are shown in Table 1. The ΔU data listed in Table 1 show that **CX** lies 0.9 kcal/

Table 3. Classical Dissociation Energy (D_e , in kcal/mol), Dissociation Energy at 0 K (D_0 in kcal/mol), and Geometrical Parameters^a (distances in Å and angles in deg) Calculated at Different CASPT2 Levels with the aug-cc-pVTZ Basis Set for *trans*-HOOO^{*} (X^2A'')

method	D_e	D_0	$r(O_aO_b)$	$r(O_bO_c)$	$r(O_cH)$	$\theta(O_aO_bO_c)$	$\theta(O_bO_cH)$	$\tau(O_aO_bO_cH)$
CASPT2(13,11)	5.4	2.6 ^b	1.214	1.734	0.973	110.7	95.2	180.0
CASPT2(19,15)	5.8	3.0 ^b	1.221	1.682	0.971	110.2	95.8	180.0
exp ^c		2.9 ± 0.1	1.225	1.688	0.972	111.0	90.0	180.0

^a Geometrical parameters defined as in Figure 1. ^b ZPVE-correction evaluated from the unscaled CASSCF(19,15)/aug-cc-pVTZ harmonic vibrational frequencies. ^c Experimental geometrical parameters from ref 24 and experimental D_0 from ref 52.

mol above the energy of *trans*-HOOO^{*} and 0.2 kcal/mol below the energy of the isolated dissociation products HO^{*} ($X^2\Pi$) and O₂ ($X^3\Sigma_g^-$). Inclusion of ZPVE corrections to energy changes the relative energy ordering of these stationary points. Thus, according to the calculated ΔE_0 values, **CX** is found to be 1.4 kcal/mol below the energy of *trans*-HOOO^{*} but 0.3 kcal/mol above the energy of the dissociation products. These results lessen the importance of the possible existence of a loosely bound complex **CX** in the exit channel of the unimolecular dissociation of *trans*-HOOO^{*} (X^2A'') to HO^{*} ($X^2\Pi$) + O₂ ($X^3\Sigma_g^-$).

We note that the present CASSCF calculations predict that the unimolecular dissociation of *trans*-HOOO^{*} (X^2A'') to HO^{*} ($X^2\Pi$) + O₂ ($X^3\Sigma_g^-$) proceeds through an exit channel involving a low energy barrier. This prediction is at variance with the experimental findings of Le Picard et al.,⁵² indicating that the inverse reaction, namely, the association of HO^{*} and O₂ yielding HOOO^{*}, occurs over a minimum potential energy path with no barrier between the reactants HO^{*} + O₂ and the product HOOO^{*}.

2.2. CASPT2 Approach. The calculations reported in this section were performed within the framework of the second-order multiconfigurational perturbation theory based on a CASSCF reference wave function (CASPT2)⁵⁹ employing an IPEA shift⁶⁰ of 0.25 au, as implemented in the MOLCAS-7.4 suite of programs,⁶¹ in conjunction with the aug-cc-pVTZ basis set. Two CASs of different sizes were used in the CASSCF reference function of the CASPT2 calculations. First, a reduced valence electron CASSCF reference wave function for *trans*-HOOO^{*} was generated by distributing the valence electrons excluding six inner electrons (seeded in the 2s atomic orbitals of the three oxygen atoms) among the active orbitals determined by using the above procedure. This indicated a CAS of 11 orbitals (7 a' and 4 a'') and led to a reference CASSCF wave function, designated by CASSCF-(13,11), formed as a linear combination of 76 230 doublet spin-adapted CSFs. For each dissociation product, the small reference CASSCF wave function was generated by distributing the corresponding valence electrons (excluding the electrons seeded in the 2s atomic orbitals of the oxygen atoms) among the active orbitals generated according to the aforementioned procedure, which resulted in five active orbitals for HO^{*} ($X^2\Pi$) and six active orbitals for O₂ ($X^3\Sigma_g^-$). This led to the CASSCF reference wave functions denoted by CASSCF(5,5) and CASSCF(8,6), respectively. Second, the all valence electron CASSCF(19,15), CASSCF-(7,7), and CASSCF(12,8) wave functions described above for *trans*-HOOO^{*} (X^2A''), HO^{*} ($X^2\Pi$), and O₂ ($X^3\Sigma_g^-$), respectively, were taken as the larger CASSCF reference wave functions employed in the CASPT2 calculations. For

Table 4. A Comparison of Classical Dissociation Energies (D_e , in kcal/mol), Dissociation Energies at 0 K (D_0 , in kcal/mol), and Central O–O Bond Lengths ($r(O_bO_c)$, in Å) Calculated for HOOO^{*}

method	D_e	D_0	$r(O_bO_c)$	ref
MCHF	13.8		1.472	27
G2M2(RCC)		1.3	1.543	30
MRMP2//CASSCF	6.0	2.8	1.750	11
QCISD(T)-CBS	5.3		1.495	35
CCSD(T)-CBS(W1U)	3.4	0.1	1.544	8
MRCIQ+Q	3.9		1.677	24
CCSD(T)//QCISD	0.1	-3.9	1.522	41
MRCI+Q	5.4	1.4	1.647	41
MRCI+Q//CASSCF	3.5		1.544	42
HCTH	9.9	6.2	1.610	42
CCSD(T)	5.2	2.5	1.589	47
CASPT2(13,11)	5.4	2.6	1.734	this work
CASPT2(19,15)	5.8	3.0	1.682	this work
exp		2.9 ± 0.1	1.688	24, ^a 52 ^b

^a Experimental $r(O_bO_c)$. ^b Experimental D_0 .

the sake of brevity, the different CASPT2 levels of calculation are designated by CASPT2(m,n), where m is the number of electrons and n the number of orbitals of the CASSCF reference wave function.

The geometries were optimized using gradients determined through a finite difference of CASPT2 energies. The values of the classical dissociation energy (designated by D_e) and D_0 , along with the geometrical parameters, calculated for *trans*-HOOO^{*} (X^2A'') at the CASPT2(13,11) and CASPT(19,15) levels are summarized in Table 3. The unscaled CASSCF-(19,15) harmonic vibrational frequencies were used to compute the ZPVE corrections to obtain D_0 from D_e . For the purpose of comparison, the values of D_e , D_0 , and $r(O_bO_c)$ calculated for HOOO^{*} by several previous theoretical studies are collected in Table 4.

The optimized geometrical parameters of the equilibrium structure computed at the CASPT2(13,11) level for *trans*-HOOO^{*} are in relatively good agreement with the values determined from the experimental rotational constants in conjunction with MRCI calculations.²⁴ In the case of the HO–OO^{*} bond, the $r(O_bO_c)$ of 1.734 Å appears to be 0.046 Å longer than the experimental value of 1.688 Å.²⁴ Fortunately, the most expensive calculation reported here (i.e., the full CASPT2(19,15) geometry optimization) predicts an $r(O_bO_c)$ of 1.682 Å, in excellent agreement with the experimental value. For the rest of the geometrical parameters, the values calculated at the CASPT2(19,15) level are similar to those obtained at the less expensive CASPT2(13,11) level. Concerning the dissociation energies, it appears (see Table 3) that the values of D_e and D_0 calculated at the CASPT2(19,15) level are 0.4 kcal/mol larger than those

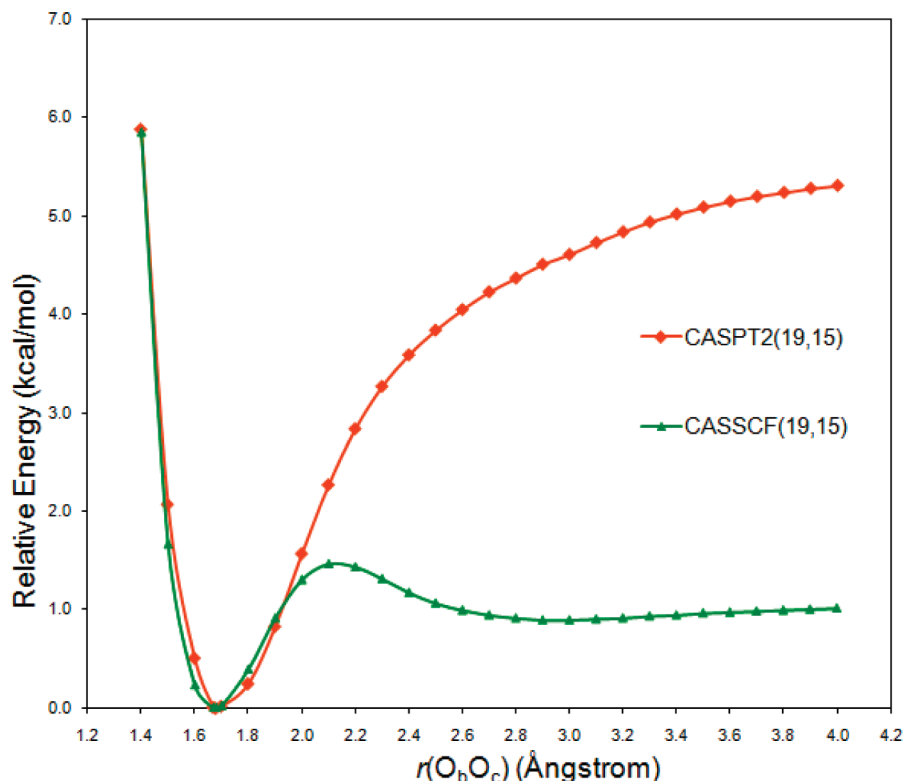


Figure 2. Potential energy profiles for dissociation of *trans*-HOOO* (X^2A'') to HO* ($X^2\Pi$) + O₂ ($X^3\Sigma_g^-$) calculated at two different levels of theory along the $r(O_bO_c)$ coordinate using geometries optimized at the CASSCF(19,15)/aug-cc-pVTZ level.

computed at the CASPT2(13,11) level. Furthermore, we note that the $D_e = 5.8$ and $D_0 = 3.0$ kcal/mol predicted by our CASPT2(19,15) calculations are about 0.5 kcal/mol higher than the theoretical $D_e = 5.2$ and $D_0 = 2.5$ kcal/mol values obtained recently by Varner et al.⁴⁷ using high-level coupled-cluster methods. In comparison with experimental values, though the CASPT2(19,15) computed $D_0 = 3.0$ kcal/mol is approximately 2.3 kcal/mol below the D_0 upper bound of 5.31 kcal/mol,⁶² it is gratifying to note that our theoretical value is in excellent agreement with the most recent experimentally determined $D_0 = 2.9 \pm 0.1$ kcal/mol.⁵²

To investigate the minimum energy reaction path of *trans*-HOOO* (X^2A'') dissociation into HO* ($X^2\Pi$) + O₂ ($X^3\Sigma_g^-$), potential energy profiles were calculated along the $r(O_bO_c)$ coordinate at different levels of theory. First, the geometries were optimized at the CASSCF(19,15)/aug-cc-pVTZ level for all degrees of freedom except a fixed $r(O_bO_c)$ distance that increased at 0.1 Å intervals. In addition, CASPT2(19,15)/aug-cc-pVTZ single-point energy calculations were performed for the geometries optimized at the CASSCF(19,15) level of theory. The resulting graphs of the potential energy profiles are shown in Figure 2. We note that at both levels of theory a minimum exists on the ground-state dissociation profile of *trans*-HOOO* (X^2A''). Starting from the minimum, at the CASSCF(19,15) level of theory, the *trans*-HOOO* (X^2A'') surmounts a very low energy barrier (ca. 1.5 kcal/mol) at $r(O_bO_c) = 2.1$ Å, which corresponds to the transition structure **TS** described above. A shallow potential energy well is observed at $r(O_bO_c) = 2.9$ Å, which corresponds to the loosely bound complex **CX** described above. In clear contrast, the graph of the

dissociation profile calculated at the CASPT2(19,15) level of theory shows that the potential energy of *trans*-HOOO* (X^2A'') increases smoothly until it reaches asymptotically the dissociation limit HO* ($X^2\Pi$) + O₂ ($X^3\Sigma_g^-$) without surmounting any energy barrier other than that imposed by the endoergicity of the dissociation.

Second, the geometries were optimized at 0.1 Å intervals for the $r(O_bO_c)$ distance at the CASPT2(13,11)/aug-cc-pVTZ level. The reoptimization of these geometries at the CASPT2(19,15)/aug-cc-pVTZ level was not attempted in the present study due to the prohibitive computational cost involved.⁶³ Nevertheless, as noted above, except for the $r(O_bO_c)$ distance, the geometrical parameters optimized at the CASPT2(19,15) level for *trans*-HOOO* (X^2A'') are close to those obtained at the less expensive CASPT2(13,11) level. Therefore, CASPT2(19,15)/aug-cc-pVTZ single-point energy calculations were performed for the geometries optimized at the CASPT2(13,11) level of theory. The resulting graphs of the potential energy profiles are shown in Figure 3. Again, the graphs of the dissociation profile calculated at either the CASPT2(13,11) or CASPT2(19,15) level of theory show that the potential energy of *trans*-HOOO* (X^2A'') increases smoothly until it reaches asymptotically the dissociation limit HO* ($X^2\Pi$) + O₂ ($X^3\Sigma_g^-$) without surmounting any energy barrier. Consequently, according to the more reliable multireference CASPT2 method, no reverse activation energy is involved in this dissociation. This prediction is in accordance with the experimental findings of Le Picard et al.,⁵² indicating that the reaction of HO* with O₂ yielding HOOO* is a barrierless association process.

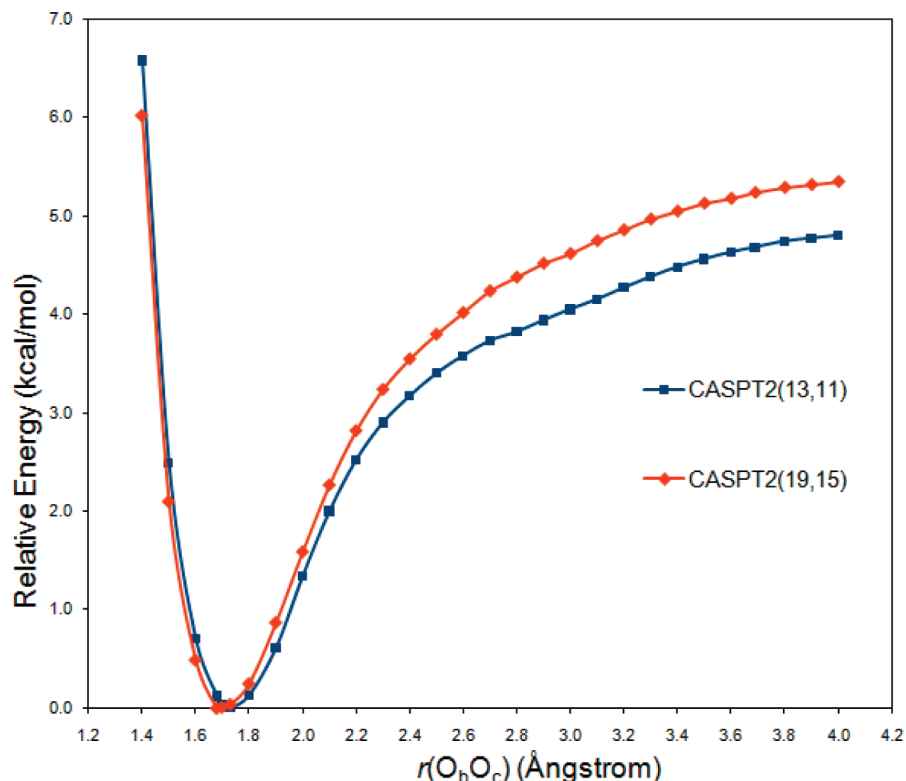


Figure 3. Potential energy profiles for dissociation of *trans*-HOOO* (X^2A'') to HO* ($X^2\Pi$) + O₂ ($X^3\Sigma_g^-$) calculated at two different CASPT2 levels along the $r(\text{O}_b\text{O}_c)$ coordinate using geometries optimized at the CASPT2(13,11)/aug-cc-pVTZ level.

3. Summary and Conclusions

Quantum-mechanical multireference-based methods CASSCF and CASPT2 were employed with the aug-cc-pVTZ basis set to investigate the equilibrium geometry of the electronic ground state of *trans*-HOOO* ($^2A''$) and its unimolecular dissociation into HO* ($X^2\Pi$) and O₂ ($X^3\Sigma_g^-$). From the analysis of the results, the following main points emerge.

The CASSCF(19,15) calculations predict for *trans*-HOOO* (X^2A'') an equilibrium structure with a HO–OO* bond length of 1.674 Å, which is in reasonable agreement with the experimental value of 1.688 Å, and give a small classical dissociation energy $D_e = 1.1$ kcal/mol. Inclusion of the ZPVE corrections leads to a 0 K dissociation energy $D_0 = -1.7$ kcal/mol, which is grossly in disagreement with the recent experimentally determined $D_0 = 2.9 \pm 0.1$ kcal/mol by Le Picard et al.

At the CASSCF(19,15) level of theory, it is found that the dissociation of *trans*-HOOO* (X^2A'') into HO* ($X^2\Pi$) and O₂ ($X^3\Sigma_g^-$) proceeds through a transition structure involving a low energy barrier of 1.5 kcal/mol. This theoretical prediction is at variance with the experimental findings of Le Picard et al., indicating that the inverse reaction, namely, the association of HO* and O₂ yielding HOOO*, occurs over a minimum potential energy path with no barrier between the reactants HO* + O₂ and the product HOOO*.

The most reliable calculation performed in this work, a full geometry optimization of *trans*-HOOO* (X^2A'') at the CASPT2(19,15)/aug-cc-pVTZ level, gives a HO–OO* bond length of 1.682 Å, which is in excellent agreement with the experimental value. At this level of theory, D_e is calculated

to be 5.8 kcal/mol. Inclusion of the zero-point energy correction (determined from CASSCF(19,15)/aug-cc-pVTZ harmonic vibrational frequencies) to this D_e leads to $D_0 = 3.0$ kcal/mol. This value of D_0 is in excellent agreement with the recent experimentally determined $D_0 = 2.9 \pm 0.1$ kcal/mol by Le Picard et al.

At the CASPT2(13,11) and CASPT2(19,15) levels of theory, we did not find for the dissociation of *trans*-HOOO* (X^2A'') a barrier other than that imposed by the endoergicity of the reaction. Consequently, no reverse activation energy is involved in this dissociation. This prediction is in accordance with the experimental findings of Le Picard et al., indicating that the reaction of HO* with O₂ yielding HOOO* is a barrierless association process.

Overall, it can be concluded that the quantum-mechanical multireference-based method CASPT2 correctly describes both the electronic structure of *trans*-HOOO* ($^2A''$) and its unimolecular dissociation into HO* ($X^2\Pi$) and O₂ ($X^3\Sigma_g^-$).

Acknowledgment. This research was supported by the Spanish MICINN (Grant CTQ2008-06536/BQU). Additional support came from the Catalanian AGAUR (Grant 2009SGR01472). The larger calculations described in this work were performed at the Centre de Supercomputació de Catalunya (CESCA).

Supporting Information Available: The total energies, zero-point vibrational energies, harmonic vibrational frequencies, and Cartesian coordinates of all structures reported in this paper. This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- (1) Murray, C.; Derro, E. L.; Sechler, T. D.; Lester, M. I. *J. Phys. Chem. A* **2007**, *111*, 4727.
- (2) Derro, E. L.; Murray, C.; Sechler, T. D.; Lester, M. I. *J. Phys. Chem. A* **2007**, *111*, 11592.
- (3) Chalmet, S.; Ruiz-Lopez, M. F. *J. Chem. Phys.* **2006**, *124*, 194502.
- (4) Chin, G. *Science* **2003**, *302*, 535.
- (5) Cooper, P. D.; Moore, M. H.; Hudson, R. L. *J. Phys. Chem. A* **2006**, *110*, 7985.
- (6) Yu, H.-G.; Varandas, A. J. C. *J. Chem. Soc., Faraday Trans.* **1997**, *93*, 2651.
- (7) Szichman, H.; Varandas, A. J. C. *J. Phys. Chem. A* **1999**, *103*, 1967.
- (8) Fabian, W. N. F.; Kalcher, J.; Janoschek, R. *Theor. Chem. Acc.* **2005**, *114*, 182.
- (9) Varandas, A. J. C. *J. Phys. Chem. A* **2004**, *108*, 758.
- (10) Yang, J.; Li, Q. S.; Zhang, S.-W. *Phys. Chem. Chem. Phys.* **2007**, *9*, 466.
- (11) Setokuchi, O.; Sato, M.; Matuzawa, S. *J. Phys. Chem. A* **2000**, *104*, 3204.
- (12) Cerkovnik, J.; Erzen, E.; Koller, J.; Pleniscar, B. *J. Am. Chem. Soc.* **2002**, *124*, 404.
- (13) Le Crane, J. P.; Rayez, J. C.; Villanave, E. *Phys. Chem. Chem. Phys.* **2006**, *8*, 2163.
- (14) Srinivasan, N. K.; Su, M. C.; Sutherland, J. V.; Michael, J. V. *J. Phys. Chem. A* **2005**, *109*, 7902.
- (15) Wentworth, P.; Jones, L. H.; Wentworth, A. D.; Zhu, X. Y.; Larsen, N. A.; Wilson, I. A.; Xu, X.; Goddard, W. A.; Janda, K. D.; Eschenmoser, A. *Science* **2001**, *293*, 1806.
- (16) Pleniscar, B.; Cerkovnik, J.; Takavec, T.; Koller, J. *Chem.—Eur. J.* **2000**, *6*, 809.
- (17) Pleniscar, B. *Acta Chim. Slov.* **2005**, *52*, 1.
- (18) Engdahl, A.; Nelander, B. *Science* **2002**, *295*, 482.
- (19) Xu, X.; Goddard, W. A. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *96*, 15308.
- (20) Alosio, S.; Francisco, J. S. *J. Am. Chem. Soc.* **1999**, *121*, 8592.
- (21) Kraka, E.; Cremer, D.; Koller, J.; Pleniscar, B. *J. Am. Chem. Soc.* **2002**, *124*, 8462.
- (22) Chalmet, S.; Ruiz-Lopez, M. F. *ChemPhysChem* **2006**, *7*, 463.
- (23) Wu, A.; Cremer, D.; Pleniscar, B. *J. Am. Chem. Soc.* **2003**, *125*, 9395.
- (24) Suma, K.; Sumiyoshi, Y.; Endo, Y. *Science* **2005**, *308*, 1885.
- (25) Blint, R. J.; Newton, M. D. *J. Phys. Chem.* **1973**, *59*, 6220.
- (26) Mathisen, K. B.; Siegbahn, P. E. M. *Chem. Phys.* **1984**, *90*, 225.
- (27) Dupuis, M.; Fitzgerald, G.; Hammond, B.; Leister, W. A.; Schaefer, H. F., III. *J. Chem. Phys.* **1986**, *84*, 2691.
- (28) Vincent, M. A.; Hillier, I. H.; Burton, N. A. *Chem. Phys. Lett.* **1995**, *233*, 111.
- (29) Speranza, M. *Inorg. Chem.* **1996**, *35*, 6140.
- (30) Jungkamp, T. P. W.; Seinfeld, J. H. *Chem. Phys. Lett.* **1996**, *257*, 15.
- (31) Speranza, M. *J. Phys. Chem. A* **1998**, *102*, 7535.
- (32) Cacace, F.; de Petris, G.; Pepi, F.; Troiani, A. *Science* **1999**, *285*, 81.
- (33) Hollebeek, T.; Ho, T. S.; Rabitz, H. *Annu. Rev. Phys. Chem.* **1999**, *50*, 537.
- (34) Nelander, B.; Engdahl, A.; Svensson, T. *Chem. Phys. Lett.* **2000**, *332*, 403.
- (35) Yu, H.-G.; Varandas, A. J. C. *Chem. Phys. Lett.* **2001**, *334*, 173.
- (36) Pei, K. M.; Zhang, X. Y.; Kong, X. L.; Li, H. Y. *Chin. J. Chem. Phys.* **2002**, *15*, 263.
- (37) Denis, P. A.; Kieninger, M.; Ventura, O. N.; Cachau, R. E.; Diercksen, G. H. F. *Chem. Phys. Lett.* **2002**, *365*, 440. Erratum **2003**, *377*, 483.
- (38) Varandas, A. C. J. *Chem. Phys. Chem.* **2005**, *3*, 453.
- (39) Janoschek, R.; Fabian, W. M. F. *J. Mol. Struct.* **2006**, *780*, 80.
- (40) Xu, Z. F.; Lin, M. C. *Chem. Phys. Lett.* **2007**, *440*, 12.
- (41) Mansergas, A.; Anglada, J. M.; Olivella, S.; Ruiz-Lopez, M. F.; Martins-Costa, M. *Phys. Chem. Chem. Phys.* **2007**, *9*, 5865.
- (42) Braams, B. J.; Yu, H.-G. *Phys. Chem. Chem. Phys.* **2008**, *10*, 3150.
- (43) Varner, M. E.; Harding, M. E.; Gauss, J.; Stanton, J. F. *Chem. Phys.* **2008**, *346*, 53.
- (44) Semes'ko, D. G.; Khursan, S. L. *Russ. J. Phys. Chem. A* **2008**, *82*, 1277.
- (45) Denis, P. A.; Ornellas, F. R. *Chem. Phys. Lett.* **2008**, *464*, 150.
- (46) Denis, P. A.; Ornellas, F. R. *J. Phys. Chem. A* **2009**, *113*, 499.
- (47) Varner, M. E.; Harding, M. E.; Vazquez, J.; Gauss, J.; Stanton, J. F. *J. Phys. Chem. A* **2009**, *113*, 11238.
- (48) Grant, D. J.; Dixon, D. A.; Francisco, J. S.; Feller, D.; Peterson, K. A. *J. Phys. Chem. A* **2009**, *113*, 11343.
- (49) Murray, C.; Derro, E. L.; Sechler, T.-D.; Lester, M. I. *Acc. Chem. Res.* **2009**, *42*, 419.
- (50) Derro, E. L.; Sechler, T. D.; Murray, C.; Lester, M. I. *J. Phys. Chem. A* **2008**, *112*, 9269.
- (51) Derro, E. L.; Sechler, T. D.; Murray, C.; Lester, M. I. *J. Chem. Phys.* **2008**, *128*, 244313.
- (52) Le Picard, S. D.; Tizniti, M.; Canosa, A.; Sims, I. R.; Smith, I. W. M. *Science* **2010**, *328*, 1258.
- (53) For a review, see: Roos, B. O. *Adv. Chem. Phys.* **1987**, *69*, 399.
- (54) Kendall, R. A.; Dunning, T. H., Jr.; Harrison, R. J. *J. Chem. Phys.* **1992**, *96*, 6796.
- (55) Anglada, J. M.; Bofill, J. M. *Theor. Chim. Acta* **1995**, *92*, 369.
- (56) (a) Schlegel, H. B. *J. Comput. Chem.* **1982**, *3*, 214. (b) Bofill, J. M. *J. Comput. Chem.* **1994**, *15*, 1.
- (57) Schmidt, M. W.; Baldrige, K. K.; Boatz, J. A.; Elbert, S. T.; Gordon, M. S.; Jensen, J.; Koseki, S.; Matsunaga, N.; Nguyen, K. A.; Su, S.; Windus, T. L.; Dupuis, M.; Montgomery, J. A. *J. Comput. Chem.* **1993**, *14*, 1347.

- (58) (a) Gonzalez, C.; Schlegel, H. B. *J. Chem. Phys.* **1989**, *90*, 2154. (b) Gonzalez, C.; Schlegel, H. B. *J. Phys. Chem.* **1990**, *94*, 5523.
- (59) (a) Anderson, K.; Malmqvist, P.-A.; Roos, B. O.; Sadlej, A. J.; Wolinski, K. *J. Phys. Chem.* **1990**, *94*, 5483. (b) Anderson, K.; Malmqvist, P.-A.; Roos, B. O. *J. Chem. Phys.* **1992**, *96*, 1218.
- (60) Ghigo, G.; Roos, B. O.; Malmqvist, P.-Å. *Chem. Phys. Lett.* **2004**, *396*, 142.
- (61) Karlström, G.; Lindh, R.; Malmqvist, P.-Å.; Roos, B. O.; Ryde, U.; Veryazov, V.; Widmark, P.-O.; Cossi, M.; Schimmelpfennig, B.; Neogrády, P.; Seijo, L. *Comput. Mater. Sci.* **2003**, *28*, 222.
- (62) At this point, it is worth pointing out that this upper bound value of D_0 was determined for the DOOO* isotopomer.⁵⁰ Derro et al.⁵¹ suggested that a reduced value of D_0 is expected

for HOOO* because the ZPVEs of HOOO* and HO* are likely to be larger than those of DOOO* and DO*, respectively. In fact, we have calculated the ZPVEs of DOOO* and DO* at the CASSCF(19,15)/aug-cc-pVTZ level of theory (see Table S2, Supporting Information). The resulting values, in conjunction with the ZPVE of O₂, give a $\Delta(\text{ZPVE})$ of -2.3 kcal/mol for the dissociation of DOOO*, which is 0.5 kcal/mol larger than the value of -2.8 kcal/mol calculated at the same level of theory for the dissociation of HOOO*. The estimated *corrected* experimental D_0 upper bound value of HOOO* is, thus, $5.3 - 0.5$ kcal/mol, namely, 4.8 kcal/mol.

- (63) It takes about 400 h of CPU time to perform a single point calculation of the energy plus energy gradients at the CASPT2(19,15)/aug-cc-pVTZ level on a 3.0 GHz Intel Xeon E5472 CPU.

CT100358E

How Well Can Kohn–Sham DFT Describe the HO₂ + O₃ Reaction?

Luís P. Viegas, Adriana Branco, and António J. C. Varandas*

Departamento de Química, Universidade de Coimbra, 3004-535 Coimbra, Portugal

Received June 29, 2010

Abstract: In a previous work (*J. Chem. Theory Comput.* 2010, 6, 412) we reported an ab initio investigation of the reaction between ozone and the hydroperoxyl radical. The studies on this atmospheric reaction are here continued with an evaluation of different exchange-correlation functionals (all rungs of “Jacob’s ladder” of density functional approximations are represented) in Kohn–Sham DFT calculations. We focus on the comparison between the barrier heights of the oxygen- and hydrogen-abstraction mechanisms here calculated with the ones previously obtained at the CASPT2(11,11)/AVTZ level. The comparison is also extended to the remaining stationary points. Additionally, a relation between the fraction of exact exchange of one-parameter hybrid functionals and the imaginary frequency of a saddle point is developed, originating three new functionals that are also used in the present benchmark calculations.

1. Introduction

Density functional theory (DFT)¹ is nowadays one of the most (if not the most) used methods of performing electronic structure calculations in the ground state of atoms, molecules, and solids. Its success stems from the simplification of the Schrödinger equation through the Hohenberg–Kohn (HK) theorems² and also from the practical implementation of the self-consistent Kohn–Sham (KS) equations.³ These look like (and scale like) the Hartree–Fock equations, where the several terms that make up the ground-state energy (which HK proved to be a functional of the electron density, $E_0[\rho]$) can be calculated exactly except one, corresponding to a small fraction of the total energy and named exchange-correlation (xc) energy functional, $E_{xc}[\rho]$, which is unknown. KS-DFT is therefore exact in principle, while in practice $E_{xc}[\rho]$ must be approximated, thus being the main source of error in the theory. Such approximate functionals are often constructed⁴ by constraint satisfaction (nonempirical functionals) or by fitting them to experimental and/or ab initio data (semiempirical functionals). It is believed that increasing the number of satisfied constraints is a step toward the exact and universal functional, but while this approach seems theoretically attractive, its progress has shown to be somewhat slow.⁵ On the other hand, semiempirical functionals rapidly achieved widespread success, particularly through the

popular B3LYP functional.⁶ The semiempirical approach has the advantage of making accurate predictions for systems which belong (or are similar) to the training set, but it carries two main problems: one is the possible failure for systems outside the training set, and the other is that the functionals sometimes do not respect some of the known exact constraints. However, their low computational cost together with a huge predictive character for systems that cannot be correctly described by nonempirical functionals explains the great success of the semiempirical approach.

While in ab initio theory one knows exactly how to proceed to improve the quality of the results, in KS-DFT this is not so obvious and straightforward. One way to hierarchize and develop improved exchange-correlation functionals is by adding to them increasingly complex ingredients, therefore creating the possibility of satisfying more constraints. This is the basic philosophy behind the “Jacob’s ladder”^{5,7} of density functional approximations to the exchange-correlation energy, where functionals (nonempirical or semiempirical) are assigned to different rungs of the ladder, according to the complexity of their ingredients. As one naturally expects, on going up the ladder, accuracy and computational cost will generally increase. The first rung is the local spin density approximation (LSDA),³ often referred to as the “mother of all approximations”.⁷ It uses as ingredients the spin densities $\rho_\alpha(\mathbf{r})$ and $\rho_\beta(\mathbf{r})$ and is by construction exact for uniform densities or densities that vary

* Corresponding author e-mail: varandas@qtvs1.qui.uc.pt.

very slowly over space. Many electronic systems do not respect these conditions (e.g., atoms and molecules), making LSDA more useful in solids. However, despite overestimating atomization and binding energies, LSDA gives good results in predicting properties like molecular geometries and vibrational frequencies,^{8–11} being a useful structural tool except for thermochemistry.¹² The second rung is the generalized gradient approximation (GGA) which introduces the density gradients $\nabla\rho_\alpha(\mathbf{r})$ and $\nabla\rho_\beta(\mathbf{r})$ as additional ingredients. GGAs show a good improvement for thermochemistry^{13,14} relative to LSDA. The third rung is the meta-generalized gradient approximation (meta-GGA), with $\nabla^2\rho_\alpha(\mathbf{r})$ and $\nabla^2\rho_\beta(\mathbf{r})$ being additional ingredients or, more commonly, the Kohn–Sham orbital kinetic energy densities $\tau_\alpha(\mathbf{r})$ and $\tau_\beta(\mathbf{r})$. The meta-GGAs mainly improve atomization energies^{7,15} while keeping computational cost similar to the previous rungs.

The fourth rung is called hyper-GGA and employs the exact exchange (Hartree–Fock-like) energy density as an ingredient. These functionals are based on the adiabatic connection method (ACM)^{12,16–18} and are also called hybrid functionals due to mixing of a fraction (a_0) of the exact exchange with GGA or meta-GGA exchange. The first hybrids were introduced by Becke,^{19,20} and as many others that followed, they use a_0 as a constant (global hybrids), most often fitted to reproduce training sets and sporadically calculated theoretically.^{19,21} However, besides not being system independent, a_0 is also not geometry independent for a given species,^{22,23} a fact that led to the so-called local hybrids,^{24–29} where $a_0(\mathbf{r})$ is a function of the coordinate space, mixing exact and DFT energy densities at each point in space. Hybrid functionals turned out to be a huge success in chemistry because of the great improvement over previous rungs in thermochemical calculations, particularly in calculating barrier heights, where exchange–correlation functionals are known to exhibit problems³⁰ (namely, by underestimating transition-state energies^{31,32}). Such problems are known to be caused in part by the self-interaction error (SIE)^{1,33} that becomes particularly important in nonequilibrium structures with stretched bonds.³⁴ Thus, they are a consequence of not including corrections^{34,35} to the SIE in the approximate exchange–correlation functionals; the use of exact exchange will then be expected to reduce the SIE in KS-DFT calculations. The fifth and final rung of the ladder utilizes not only the occupied KS orbitals but also the unoccupied ones.⁵ A special case of fifth-rung functionals are the double hybrids,³⁶ as they mix wave function methods with hybrid DFT ones. Typically, in a double-hybrid calculation, the KS orbitals and eigenvalues resulting from the self-consistent run are subsequently used in a MP2-like calculation of the correlation energy, which replaces some of the GGA or meta-GGA correlation.

The atmospherically important $\text{HO}_2 + \text{O}_3$ reaction^{37–42} has been the subject of recent theoretical investigations,^{43–46} where extensive ab initio calculations were performed in order to clarify its mechanism. Special emphasis was given to the saddle points that represent the attack of the hydroperoxyl radical to ozone from the O and H sides (SP_1 and SP_4 of ref 46, respectively), since these are the critical regions

of the potential-energy surface (PES) that determine the dynamics of the reaction. Our goal with the present study is a fundamental one, and it is reflected in the article’s title. We will answer the posed question by assessing the quality of several exchange–correlation functionals (belonging to different rungs of the “Jacob’s ladder”) in the description of the reaction mechanism. Such a methodology has been suggested in ref 47, where it is stated “Users should also report results on several different rungs, where possible, both as a check on consistency and as a guidance for functional developers.” Again, our focus will be on calculation of the SP_1 and SP_4 barrier heights, with the best performing functionals being then used to study and compare the whole reaction mechanism here obtained with the one calculated and graphically represented in Scheme 1 of ref 46. Thus, the KS-DFT results will be compared to the ones calculated at the CASPT2(11,11)/AVTZ//CASSCF(11,11)/6-311++G(2df,2p) level of theory, which are probably the most reliable ab initio results reported thus far in the literature for the description of all stationary points of the $\text{HO}_5(^2\text{A})$ PES, although multireference perturbation theory (MRPT) may not be free from problems.^{48–52} Recall that the title reaction is a challenging one to theoretical methods,⁵³ since some of the regions of the PES show a high multireference character.⁴⁶ In addition, knowing that KS-DFT is sometimes problematic in describing saddle points makes it an interesting reaction for the benchmarking of KS-DFT. Finding a functional suitable for electronic calculations in the $\text{HO}_5(^2\text{A})$ PES would be extremely useful in the study of the $\text{HO}_2 + \text{O}_3$ reaction, since the computational cost associated with KS-DFT is much lower than multireference perturbation theory, thus allowing an extensive mapping of the PES at the regions of interest.

The structure of the paper is as follows. In Computational Methods, the basic theory and technical details behind the calculations are described, whereas the next section reports the Results and Discussion. The conclusions are gathered in the last section.

2. Computational Methods

All calculations have been performed with the GAMESS⁵⁴ and ORCA⁵⁵ packages. The GAMESS code was used for all KS-DFT optimizations employing the M06 family of exchange–correlation functionals and also for all intrinsic reaction coordinate (IRC) calculations, while ORCA was used in the remaining calculations. While using both computer codes, a vibrational analysis of the harmonic vibrational frequencies was performed after each geometry optimization to confirm the nature of the stationary points. The MacMolPlt⁵⁶ graphical user interface was used for visualization of the different geometric and electronic features of the PES.

The following DFT functionals, ranked according to the “Jacob’s ladder” of Perdew, have been considered in this work: first rung, LSDA;³ second rung, BLYP^{57,58} and PBE;⁵⁹ third rung, TPSS;¹⁵ fourth rung (hybrid GGAs), B3LYP,⁶ PBEh,^{60,61} and BH&HLYP;^{19,57,58} fourth rung (hybrid meta-

GGAs), M06,⁶² M06-2X,⁶² and M06-HF,⁶³ fifth rung, B2PLYP⁶⁴ and B2GP-PLYP.⁶⁵

The following basis sets were used: 6-311++G(2df,2p), 6-311++G(2df,2pd), and aug-cc-pVXZ ($X = D, T$ or simply AVXZ). Geometry optimizations were performed with the 6-311++G(2df,2p) basis set to allow a direct comparison between the geometrical parameters calculated here and in our previous study.⁴⁶ However, this basis set is not available in the ORCA package, and therefore, while using this code, we adopted the 6-311++G(2df,2pd) basis instead, which only adds 5 contracted basis functions in a calculation on the HO₅(²A) PES. We expect no significant changes with such a small difference between both basis sets. The energetic parameters, such as relative energies (with special emphasis on barrier heights), relaxation energies, and basis set superposition errors (BSSE), were calculated with Dunning basis sets, since the single-point energies of ref 46 were obtained with the AVTZ basis set.

Dispersion corrections to the DFT energies (DFT-D) were also included in the calculations. These are available in GAMESS and ORCA with Grimme's formulation,^{66,67} and we performed a separate set of calculations by adding this correction to the BLYP, PBE, TPSS, B3LYP, and B2PLYP exchange-correlation functionals. The details and testing of this method can be found in the original publications^{66,67} and also in some recent papers.^{68–76} We should stress that no saddle points were found with TPSS-D, and therefore, no results with this functional will be shown.

All barrier heights (except the ones involving the M06 family of functionals) were calculated with BSSE corrections according to the counterpoise (CP) scheme proposed by Boys and Bernardi.⁷⁷ The relative energies were obtained by taking into account the geometric modifications of the fragments upon formation of the saddle points^{78,79}

$$\Delta E^{\text{CP}} = E_{\text{SP}}^{\text{opt}} - \sum_{m=1}^N E_m^{\text{opt}} + \sum_{m=1}^N (E_m^{\text{frz}} - E_m^{\text{frz},*}) \quad (1)$$

where $E_{\text{SP}}^{\text{opt}}$ is the optimized energy of the saddle point, the E_m 's are the energies of the monomers (in this case we have two monomers, O₃ and HO₂, and so $N = 2$). The superscript "opt" denotes the individually optimized monomers, and "frz" refers to the monomers frozen in their saddle-point geometries. The asterisk (*) indicates a calculation with ghost orbitals. From these quantities useful information can be extracted, such as the BSSE itself (third term in the rhs of eq 1, which is expected to become smaller with increasing size of the basis set and to approach zero when the complete basis set (CBS) limit is reached⁷⁸), and the relaxation energy of each monomer, given by $E_m^{\text{frz}} - E_m^{\text{opt}}$, which gives us an indication of how much the geometry of one monomer changes until it reaches its saddle-point geometry. Equation 1 could not be applied to the M06 functionals, since GAMESS does not support BSSE runs with KS-DFT calculations. Instead, and while using GAMESS, we adopted the strategy used before, by calculating the reactants as a supermolecule with the fragments separated by 150 Å, not including CP corrections. A further remark to note is that eq 1 and CP are not free from criticism, particularly for

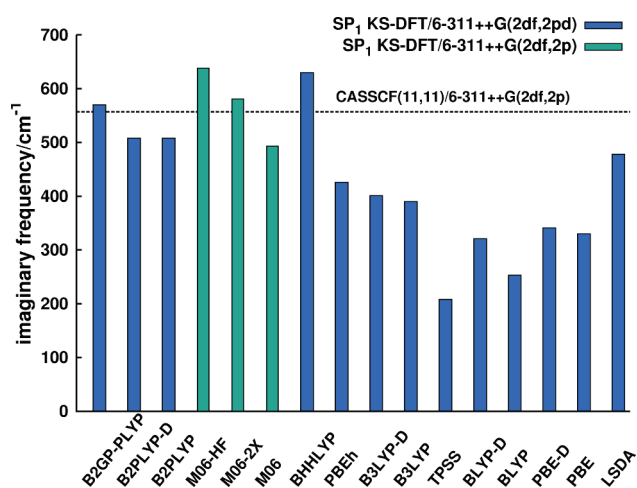


Figure 1. Comparison between the SP₁ imaginary frequency calculated at the CASSCF(11,11)/6-311++G(2df,2p) level (dashed line) and at the KS-DFT level with two different basis sets and different exchange-correlation functionals.

transition states (and other topographical features) that are at regions far from the separated reactants, thus making unclear what kind of correction is provided by considering the separate reactants at their geometries in the transition state (suffice to note that interacting fragments are not rigid blocks and that CP often overestimates the BSSE). Therefore, the recommendable approach would be CBS extrapolation.^{80,81} However, in the absence of CBS extrapolation schemes for the various components of the KS-DFT energy, CP will be employed here for necessity.

In addition, and because the B2GP-PLYP results were included at the last minute with the most recent ORCA version (2.7 revision 0), we stress that all energetic parameters calculated with this functional and with the AVTZ basis set were in fact obtained at the B2GP-PLYP/AVTZ//B2GP-PLYP/AVDZ level, i.e., we did not perform optimizations with the AVTZ basis set in order to save computational time. Judging from previous calculations during the course of this work, the associated error with this approach should be less than 0.1 kcal mol⁻¹.

3. Results and Discussion

3.1. Saddle-Point Structures. We begin by comparing the imaginary frequencies of SP₁ and SP₄ obtained here with the ones obtained in ref 46. This can be seen in Figures 1 (SP₁) and 2 (SP₄), with the color indicating different basis sets. We also distributed the exchange functionals in the X axis according to their rank in the "Jacob's ladder", starting with LSDA (first rung) on the far right and ending with B2GP-PLYP (fifth rung) on the far left. In Figure 1, we can see a correlation between the rung to which each functional belongs and the quality of the calculated imaginary frequency. The exception is LSDA, which achieves a very good result, despite being the simplest functional in the set, and TPSS, which is the worse performing functional for such an imaginary frequency. The inclusion of dispersion corrections has also shown to improve the results, especially with BLYP, where an increase of 68 cm⁻¹ is observed. Note also that B3LYP and PBEh performed quite poorly, despite the fact

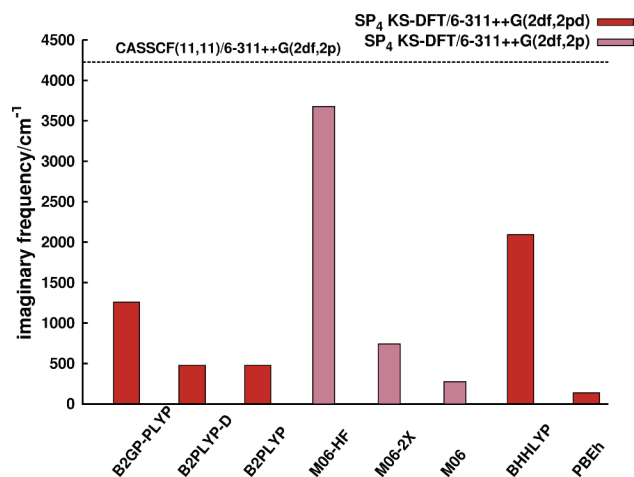


Figure 2. Comparison between the SP_4 imaginary frequency calculated at the CASSCF(11,11)/6-311++G(2df,2p) level (dashed line) and KS-DFT level with two different basis sets and different exchange-correlation functionals.

that they are fourth-rung functionals. This is most likely caused by their low percentage of exact exchange (20% and 25%, respectively).

Figure 2 shows a drastic reduction in the number of exchange functionals. This is because we were not able to optimize SP_4 using functionals without exact exchange (first three rungs). Another interesting result is that SP_4 could not be optimized with the popular B3LYP functional. In fact, the best result seen in Figure 2 is obtained with M06-HF, which incorporates 100% of exact exchange, followed by BH&HLYP with 50% of exact exchange and also B2GP-PLYP. The remaining imaginary frequencies show a huge disagreement with the CASSCF result, with no correlation between the functional's rung and the quality of its imaginary frequency. However, the role of the percentage of exact exchange reflects the crucial importance of the SIE in this saddle point, a common situation in hydrogen-abstraction mechanisms.^{82–88} In this particular case, in which a hydrogen atom is transferred to ozone, we are in the presence of an odd-electron problem (three electrons in this case), which is known to be seriously plagued with SIE (see refs 86–88 and references therein). It is therefore natural that functionals with a higher percentage of exact exchange give better results since they “need” this contribution to cancel out the SIE originating in the Coulomb electron–electron interaction. This is also the case with the B2PLYP functional (53% of exact exchange), which has a lower percentage of exact exchange than double hybrids specifically parametrized for the calculation of barrier heights,^{65,89} like B2GP-PLYP.

We also wanted to assess the quality of the saddle-point structures, so for each geometry we calculated its perpendicular looseness,⁹⁰ defined by

$$R_{\text{sum}}^{\ddagger} = R^{\ddagger}(\text{breaking bond}) + R^{\ddagger}(\text{forming bond}) \quad (2)$$

and compared it with the ab initio value by means of the equation

$$\Delta R_{\text{sum}}^{\ddagger} = R_{\text{sum}}^{\ddagger}[\text{KS-DFT}] - R_{\text{sum}}^{\ddagger}[\text{ab initio}] \quad (3)$$

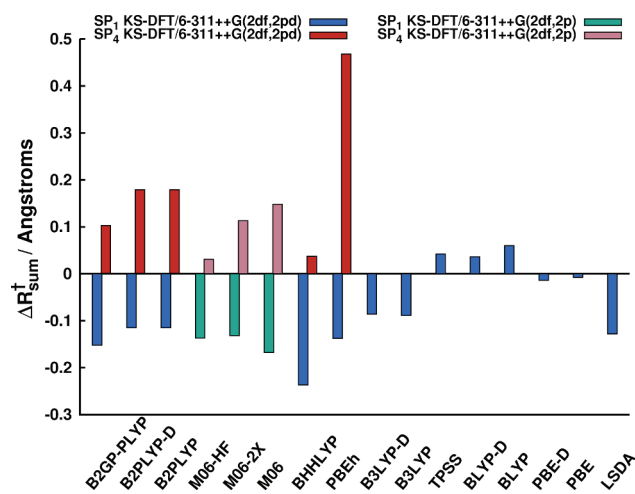


Figure 3. Comparison between the SP_1 and SP_4 perpendicular looseness calculated at the CASSCF(11,11)/6-311++G(2df,2p) level and KS-DFT level with two different basis sets and different exchange-correlation functionals (see text and eqs 2 and 3 for more details).

with the ab initio calculations performed at the CASSCF(11,11)/6-311++G(2df,2p) level.⁴⁶ The perpendicular looseness given by eq 2 “is a measure of the looseness of the structure in a direction perpendicular to the reaction coordinate”.⁹⁰ It also gives us a rough measure of the quality of the calculated geometry, especially in the important region of the saddle-point structure, where one bond is broken and another one is formed. The results can be seen in Figure 3. Interestingly, for SP_1 , the better (as before, by better we mean closest to the ab initio values) results are obtained with second- and third-rung functionals, being followed by B3LYP. This is also a known consequence of the presence of exact exchange; the molecular geometries are frequently better for functionals without it.^{62,84} The majority of these saddle points is tighter than the CASSCF structures. As for SP_4 , the best results are again obtained with BH&HLYP and M06-HF, with all of the structures being looser than the ab initio calculations. A particular bad result is obtained with PBEh, with its saddle point having a perpendicular looseness of almost 0.5 Å. In this particular case, the PBEh forming bond is ~ 0.7 Å larger than the CASSCF forming bond, being the main source of error and clearly indicating a poor quality structure.

3.2. Barrier Heights. Figure 4 shows a graphical comparison between the KS-DFT and ab initio barrier heights; a numerical, more detailed, comparison is provided later. The SP_1 barrier height calculated with the LSDA functional (-11.72 kcal mol⁻¹) is not included in the figure in order to facilitate the visualization of the remaining barrier heights. For this saddle point the best barrier heights are the ones calculated with some of the highest rung functionals, namely, B2GP-PLYP and M06-HF. The widely used B3LYP functional shows disappointing results concerning barrier heights, but this is not surprising as it is well known that this functional often underestimates barrier heights due to the lack of exact exchange. A good example is given in ref 91. The inexistence of exact exchange can also be seen to

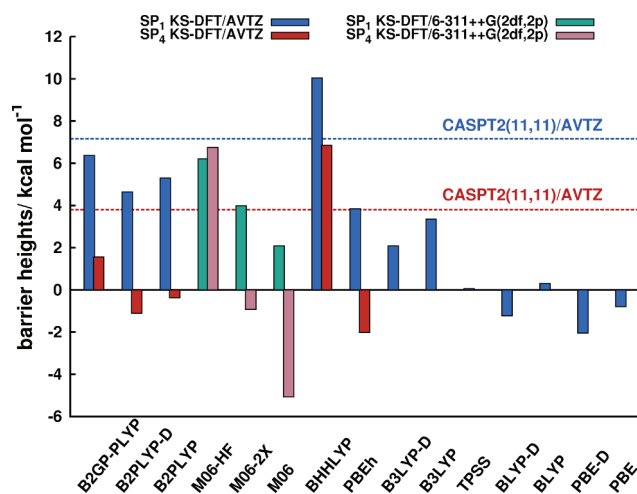


Figure 4. Comparison between the SP_1 and SP_4 barrier heights calculated at the CASPT2(11,11)/AVTZ level (SP_1 , blue dashed line; SP_4 , red dashed line) and KS-DFT level with two different basis sets and different exchange-correlation functionals.

dramatically affect the results obtained with functionals from the three lowest rungs.

The best functionals in calculating the barrier heights of SP_4 are again (as it happened for the imaginary frequencies and perpendicular looseness) B2GP-PLYP, M06-HF, and BH&HLYP. However, M06-HF calculates an SP_4 barrier higher than the SP_1 one, contrary to experimental and ab initio results. In the case of BH&HLYP the difference between the SP_1 and SP_4 barriers (~ 3 kcal mol⁻¹) is consistent with the CASPT2(11,11)/AVTZ results. The remaining functionals do a very poor job in describing the SP_4 barrier height, being energetically below the reactants. Note that PBEh (25% of exact exchange) has the second worse barrier height and shows the worse results for the imaginary frequency and perpendicular looseness. This is again explained by the large SIE present in transition states of chemical reactions, particularly hydrogen abstractions involving ozone.^{86,87} The barrier heights of transition states calculated with KS-DFT are typically too low, which might be an indication that the BH&HLYP functional carries an excess of exact exchange for this specific reaction, since both barrier heights are well above the CASPT2 results.

In Tables 1 and 2 we present the barrier heights, BSSE, and relaxation energies calculated for SP_1 and SP_4 with the different exchange-correlation functionals and using the AVDZ and AVTZ basis sets, respectively. For reasons mentioned before, the numerical results concerning the M06 functionals are not included in these tables but will appear in a subsequent section. With the exception of the LSDA functional, all barrier heights of Table 2 were used in Figure 4. Note that by subtracting the BSSE (third column of Tables 1 and 2) from ΔE^{CP} in the second column one obtains the CP-free barrier heights. Concerning the BSSE, some comments should be made at this point. First, one observes the expected rapid decrease of the BSSE as the basis set size increases.^{80,92} Excluding the double hybrids, all functionals show a BSSE below 0.4 kcal mol⁻¹ with the AVTZ basis set, which is quite an acceptable result with this moderate

Table 1. Energetic Parameters Available by Computing the Barrier Heights Using Eq 1^a

method/SP	ΔE^{CP}	BSSE	O ₃ relaxation	HO ₂ relaxation
CASPT2(11,11)/SP₁	7.160			
B2GP-PLYP/SP ₁	7.403	1.900	1.720	0.038
B2PLYP-D/SP ₁	5.333	1.488	1.489	0.011
B2PLYP/SP ₁	6.019	1.488	1.492	0.011
BH&HLYP/SP ₁	10.609	0.901	1.928	0.184
PBEh/SP ₁	4.112	0.767	1.517	0.008
B3LYP-D/SP ₁	2.343	0.603	1.671	0.011
B3LYP/SP ₁	3.563	0.535	1.647	0.005
TPSS/SP ₁	-0.309	0.475	1.575	0.069
BLYP-D/SP ₁	-1.384	0.309	1.885	0.052
BLYP/SP ₁	0.379	0.421	1.592	0.016
PBE-D/SP ₁	-2.017	0.497	1.907	0.080
PBE/SP ₁	-0.907	0.469	1.859	0.083
LSDA/SP ₁	-11.829	0.587	3.101	1.478
CASPT2(11,11)/SP₄	3.810			
B2GP-PLYP/SP ₄	2.077	1.653	0.383	2.952
B2PLYP-D/SP ₄	-0.867	1.242	0.392	1.384
B2PLYP/SP ₄	-0.128	1.240	0.393	1.385
BH&HLYP/SP ₄	6.751	0.790	0.600	5.953
PBEh/SP ₄	-2.029	0.634	0.079	0.140

^a The definition of BSSE and relaxation energies are given at the end of section 2. The energies are given in kcal mol⁻¹, and all calculations were performed with the AVDZ basis set. The CASPT2/AVTZ results were obtained from ref 46.

Table 2. Energetic Parameters Available by Computing the Barrier Heights Using Eq 1^a

method/SP	ΔE^{CP}	BSSE	O ₃ relaxation	HO ₂ relaxation
CASPT2(11,11)/SP₁	7.160			
B2GP-PLYP/SP ₁	6.370	1.107	2.078	0.088
B2PLYP-D/SP ₁	4.644	0.825	1.348	0.012
B2PLYP/SP ₁	5.309	0.825	1.351	0.012
BH&HLYP/SP ₁	10.046	0.301	1.891	0.194
PBEh/SP ₁	3.844	0.264	1.393	0.009
B3LYP-D/SP ₁	2.093	0.182	1.513	0.009
B3LYP/SP ₁	3.360	0.195	1.530	0.006
TPSS/SP ₁	0.069	0.310	1.121	0.017
BLYP-D/SP ₁	-1.226	0.268	1.964	0.053
BLYP/SP ₁	0.303	0.098	1.343	0.019
PBE-D/SP ₁	-2.038	0.133	1.657	0.069
PBE/SP ₁	-0.792	0.248	1.597	0.064
LSDA/SP ₁	-11.720	0.355	2.940	1.521
CASPT2(11,11)/SP₄	3.810			
B2GP-PLYP/SP ₄	1.558	1.099	0.443	3.000
B2PLYP-D/SP ₄	-1.113	0.810	0.416	1.585
B2PLYP/SP ₄	-0.379	0.810	0.417	1.586
BH&HLYP/SP ₄	6.860	0.275	0.649	6.304
PBEh/SP ₄	-2.024	0.184	0.053	0.104

^a The definition of BSSE and relaxation energies are given at the end of section 2. The energies are given in kcal mol⁻¹, and all calculations were performed with the AVTZ basis set. The CASPT2/AVTZ results were obtained from ref 46.

size basis set. Now we turn to the double-hybrids BSSE results. Clearly, the BSSE is now higher than before, a result that can be rationalized by acknowledging that these functionals perform an MP2-type calculation, and these are known to converge rather unsystematically to the CBS limit.⁹³ As for the relaxation energies, one observes different behaviors in O₃ and HO₂. The O₃/SP₁ relaxation energies tend to decrease with increasing size of the basis set (except for B2GP-PLYP and BLYP-D), while for SP₄ this tendency is reversed, with PBEh being the only functional for which

its relaxation energy diminishes. The HO₂ relaxation energies generally increase with increasing size of the basis set, except for B3LYP-D, TPSS, PBE(-D) (SP₁), and PBEh (SP₄). These relaxation energies indirectly tell us what happens in each reaction channel when going from reactants to these two different saddle points in the PES. In the oxygen-abstraction channel (SP₁) it is clear that the O₃ fragment suffers more drastic geometry changes than HO₂ because the ozone relaxation energy is always considerably larger than the hydroperoxyl one. This was likely to be observed, since one of the bonds in ozone is broken in this channel.^{43,45,46} A similar analysis can be made to the hydrogen-abstraction channel (SP₄), where higher relaxation energies are now observed for the hydroperoxyl radical. These are consistent with the knowledge that in this reaction channel the OH bond is broken and that the O–O bond becomes smaller in the saddle-point geometry.^{45,46} Because B2GP-PLYP and BH&HLYP give the best barrier heights for SP₄ in Tables 1 and 2, one should pay particular attention to the HO₂ relaxation energies; they are in fact higher than the other three relaxation energies for this fragment. The information obtained from analyzing the relaxation energies might be useful in the future if one wants to map the PES in the region of the saddle points.

3.3. Improving One-Parameter Hybrids. In the previous two subsections we have seen that exchange-correlation functionals studied in this work have generally shown great difficulties in describing several aspects of SP₁ and SP₄, with the latter being most problematic. By examining some relations between results obtained up to this point (imaginary frequencies, perpendicular looseness, and barrier heights) and also recognizing the tight relation between the quality of barrier heights and the fraction of exact exchange in KS-DFT, we decided to investigate further some of the involved connections. An investigation concerning the imaginary frequency seemed the most obvious choice, since it is known that its magnitude is directly associated with the height of the barriers.^{85,94} A good description of the imaginary frequency ensures that the saddle point has the correct topology, which is a step toward obtaining a better barrier height. Since we also know that exact exchange has a major role in the calculation of barrier heights, we were compelled to study the relation between the percentage of exact exchange present in a functional and a saddle point's imaginary frequency. For simplicity, only one-parameter hybrids are investigated with an exchange-correlation expression of the type

$$E_{xc} = a_0 E_x^{\text{exact}} + (1 - a_0) E_x^{\text{DFT}} + E_c^{\text{DFT}} \quad (4)$$

Our procedure was to vary the fraction of exact exchange [much in the spirit of the specific reaction parameter (SRP) method of ref 95; see also ref 96], a_0 in eq 4, and observe the behavior of the imaginary frequency. This can be done easily in ORCA, where there is the possibility of choosing the building blocks in an exchange-correlation functional. We did this by using three different forms of eq 4. The first one has $E_x^{\text{DFT}} = \text{B88}$ and $E_c^{\text{DFT}} = \text{LYP}$ (with $a_0 = 0.5$ one gets BH&HLYP), the second one has DFT = PBE (with $a_0 = 0.25$ one gets PBEh), and the third one has DFT = TPSS

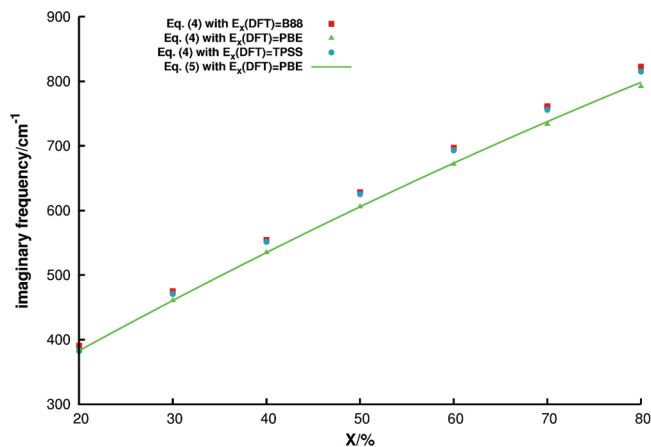


Figure 5. Imaginary frequencies of SP₁ (squares, triangles, and circles) calculated at the KS-DFT/AVDZ level using three different exchange functionals of the form of eq 4 as a function of the percentage of exact exchange ($X = 100a_0$). Also represented in this figure is the curve obtained using eq 5 with $E_x^{\text{DFT}} = \text{PBE}$ with parameters given in Table 3.

(with $a_0 = 0.10$ one gets TPSSh,⁹⁷ a hybrid meta-GGA functional). Recall that the ab initio imaginary frequency was obtained with the 6-311++G(2df,2p) basis set, so in principle we should use the closest basis set possible (in this case, the 6-311++G(2df,2pd)). However, because the differences in the imaginary frequencies using AVDZ and 6-311++G(2df,2pd) are very small (most often around 3 or 4 cm⁻¹), we performed the saddle-point optimizations with the AVDZ basis set, thus saving considerable computational time while not compromising the main conclusions drawn from this part of our study. a_0 was varied from 0.2 to 0.8, in intervals of 0.1 with each of the three forms of eq 4, thus covering a large fraction of exact exchange. The resulting set of points is represented in Figure 5, where a strong resemblance between the three sets of points can be observed, especially the ones for which $E_x^{\text{DFT}} = \text{B88}$ and TPSS. Visual inspection of Figure 5 reveals that in order to match the ab initio imaginary frequency of 557i cm⁻¹, one needs approximately 40% of exact exchange. The next step was to solve a system of linear equations in order to find the coefficients for the expression that calculates the imaginary frequency (ω^\ddagger) of SP₁ as a function of the percentage of exact exchange ($X = 100a_0$)

$$\omega^\ddagger = aX^2 + bX + c \quad (5)$$

The values of X used to solve the system of equations were naturally chosen as $X = 30, 40,$ and 50 , because they are the ones closest to our approximate prediction of X . Figure 5 shows the calculated curve based on eq 5 for which $E_x^{\text{DFT}} = \text{PBE}$. We then used the three calculated curves given by eq 5 to extract X for which $\omega^\ddagger = 557i$ cm⁻¹. The relevant data is collected in Table 3. Note that opting for a fit to the calculated points, the values of X differ from the ones given in Table 3 at most by 0.07%. The values obtained for X define three new exchange-correlation functionals: BLYP-VBV, PBE-VBV, and TPSS-VBV (the names were given to facilitate the reading of the remaining of this paper). Interestingly, our calculated values of X are very similar to

Table 3. Parameters Determined for Eq 5^a

exchange functional	a/cm^{-1}	b/cm^{-1}	c/cm^{-1}	interpolated $X/\%$
B88	-0.02560	9.7100	206.71	40.4
PBE	-0.01665	8.5895	217.83	43.1
TPSS	-0.03505	10.5255	186.35	40.7

^a The imaginary frequencies were obtained by saddle-point optimizations with the AVDZ basis set using three different exchange-correlation functionals of the form of eq 4, with each different exchange functional used being shown in the first column. The last column gives the interpolated value of X for which $\omega^\ddagger = 557i \text{ cm}^{-1}$. The new values of X define the three new functionals used in this work.

the ones used in one-parameter hybrids specially designed for calculating accurate barrier heights, such as MPW1K⁹⁸ (hybrid GGA) and BB1K⁹⁹ (hybrid meta-GGA) which incorporate 42.8% and 42% of exact exchange, respectively. Note that ORCA has the exchange and correlation functionals necessary to build the one-parameter hybrid mPW1PW¹⁰⁰ ($a_0 = 0.25$) and therefore MPW1K ($a_0 = 0.428$), but unfortunately we encountered severe numerical difficulties while using this particular implementation of the mPW exchange functional.¹⁰⁰ For this reason, we were forced to stop the calculations with this hybrid. We further note that the exchange and correlation functionals that make up PBE-VBV and MPW1K have a close similarity,⁴ which leads us to believe that because of this and the identical fractions of exact exchange they should yield very similar results.

We also tested this approach with the SP₄ imaginary frequencies, but the fraction of exact exchange needed to reproduce the ab initio imaginary frequency was very high, approximately 75%. This led to huge barrier heights, and therefore, we discarded the optimization of SP₄ using the described procedure as a valid way to calculate optimum X values. There are several possible explanations for this. One is that the ab initio imaginary frequency is miscalculated. Another is that the three tested one-parameter hybrid functionals cannot correctly calculate this frequency in such a problematic region of the PES. Although we cannot discard the former, we believe that the latter hypothesis is the correct one, since the imaginary frequency obtained with M06-HF has an acceptable agreement with the ab initio value, thus suggesting that it is possible to have a functional that calculates reasonable imaginary frequencies and barrier heights for SP₄. Ideally, application of the above procedure separately to each saddle point would lead to the same fraction of exact exchange. However, because we believe (for reasons mentioned earlier) that not all such one-parameter hybrids may be able to mimic correctly (in the sense of matching the ab initio information) the imaginary frequency at SP₄ with a moderately low value of X , we have chosen instead to employ the parameters resulting from the SP₁ analysis to predict other regions of configuration space, SP₄ included. Of course, one could always opt for a fit to the barrier heights or even to the difference between SP₁ and SP₄. This was not the followed strategy for several reasons: (a) fits to barrier heights have been massively done in the literature; (b) it would be computationally much more expensive due to the basis set sizes and number of calculations needed to accurately perform such calculations; (c) the

Table 4. Energetic Parameters of the New Functionals Defined in This Work^a

method/SP	ΔE^{CP}	$\Delta E^{\text{CP}} + \text{ZPE}$	BSSE	O ₃ relaxation	HO ₂ relaxation
CASPT2(11,11)/SP₁	7.160				
BLYP-VBV/SP ₁	8.435	10.185	0.755	1.866	0.099
PBE-VBV/SP ₁	8.307	10.147	1.033	1.645	0.069
TPSS-VBV/SP ₁	8.808	10.628	1.119	1.749	0.105
CASPT2(11,11)/SP₄	3.810				
BLYP-VBV/SP ₄	3.470	2.370	0.731	0.576	4.341
PBE-VBV/SP ₄	2.680	1.830	0.957	0.455	3.775
TPSS-VBV/SP ₄	4.126	2.996	0.979	0.490	4.596

^a The barrier heights were computed using eq 1, and the definition of BSSE and the relaxation energies are given at the end of section 2. The energies are given in kcal mol⁻¹, and all calculations were performed with the AVDZ basis set, except for the CASPT2/AVTZ results, obtained from ref 46.

Table 5. Energetic Parameters of the New Functionals Defined in This Work^a

method/SP	ΔE^{CP}	$\Delta E^{\text{CP}} + \text{ZPE}$	BSSE	O ₃ relaxation	HO ₂ relaxation
CASPT2(11,11)/SP₁	7.160				
BLYP-VBV/SP ₁	8.030	9.770	0.275	1.842	0.107
PBE-VBV/SP ₁	7.822	9.622	0.303	1.586	0.080
TPSS-VBV/SP ₁	8.453	10.253	0.370	1.721	0.127
CASPT2(11,11)/SP₄	3.810				
BLYP-VBV/SP ₄	3.565	2.415	0.242	0.624	4.664
PBE-VBV/SP ₄	2.852	1.842	0.317	0.501	4.118
TPSS-VBV/SP ₄	4.466	3.186	0.393	0.565	5.033

^a The barrier heights were computed using eq 1, and the definition of BSSE and the relaxation energies are given at the end of section 2. The energies are given in kcal mol⁻¹, and all calculations have been performed with the AVTZ basis set. The CASPT2/AVTZ results were obtained from ref 46.

adopted strategy looked more interesting because of the fewer related publications that are available.

Having determined the relation between X and ω^\ddagger , one is now faced with an obvious question: What will happen to the barrier heights when calculated with these three new functionals? To answer this question, we made new optimizations based on the BLYP-VBV, PBE-VBV, and TPSS-VBV functionals and recalculated the SP₁ and SP₄ barrier heights. These calculations were again performed with the AVDZ and AVTZ basis sets, like the ones presented in Tables 1 and 2. The new results can be seen in Tables 4 and 5 and also in Figure 6. As before, by subtracting the BSSE in the fourth column from ΔE^{CP} in the second one, one obtains the barrier heights without CP correction. The improvement of the barrier heights calculated with the new functionals can be easily acknowledged, especially for SP₄. Similar results are expected with the MPW1K and BB1K functionals because they are also one-parameter hybrids with a very similar fraction of exact exchange. Between the three new functionals, the best performance comes from BLYP-VBV and TPSS-VBV, depending on whether one is looking at the average error in the absolute barrier heights or at the error in the difference between SP₁ and SP₄, respectively. The inclusion of the zero-point energies (ZPE) of the reactants and saddle points in the calculation of the barrier heights widens the gap between SP₁ and SP₄ even further, since it lowers SP₄ and increases SP₁.

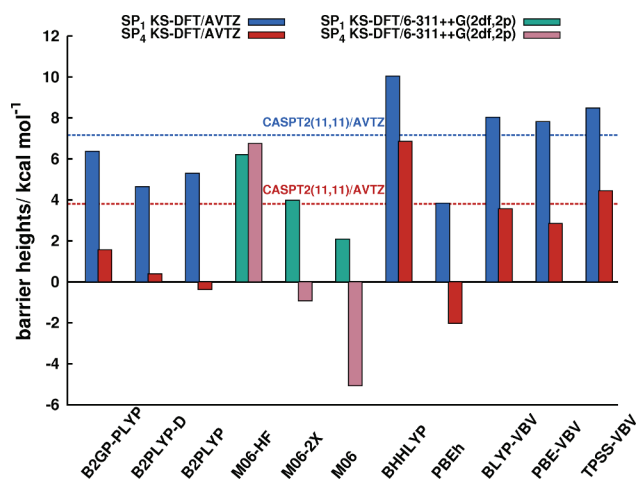


Figure 6. Comparison between the SP_1 and SP_4 barrier heights calculated at the CASPT2(11,11)/AVTZ level (SP_1 , blue dashed line; SP_4 , red dashed line) and KS-DFT level with two different basis sets and with all exchange-correlation functionals used in this work capable of optimizing both saddle points.

Table 6. Relative Energies, in kcal mol⁻¹, of the Different Stationary Points Along the IRC Path, Relative to the Reactants, Calculated with the 6-311++G(2df,2p) Basis Set^a

method	MIN ₁	MIN ₂	MIN ₃	SP ₁	SP ₂	SP ₃	SP ₄
CASPT2(11,11)	-3.50	-23.92	-35.06	7.16	-3.43	-20.88	3.81
M06-HF	-5.23	-16.53	-45.14	6.20		-14.18	6.75
M06-2X	-6.02		-43.89	3.99			-0.93
M06	-6.04		-38.16	2.09	-5.76		-5.07
BH&HLYP	-3.30		-38.53	9.72	-3.20		6.33
BLYP-VBV	-3.14		-39.12	7.74	-3.02		3.14
PBE-VBV	-3.53		-39.70	7.48			2.23
TPSS-VBV	-3.04		-38.03	7.95	-2.95		3.68

^a The single-point CASPT2(11,11)/AVTZ energies are shown for comparison.

At this point, six exchange-correlation functionals stand out from the rest: B2GP-PLYP, M06-HF, BH&HLYP, BLYP-VBV, PBE-VBV, and TPSS-VBV.

3.4. IRC Calculations. Knowing which functionals calculate better both barrier heights, we proceed by evaluating their performance in the description of the IRC path. Do these functionals calculate the same number and type of stationary points along the reaction path as the previous⁴⁶ CASSCF(11,11)/6-311++G(2df,2p) calculations? How do the relative energies compare to the CASPT2 results? Since ORCA does not perform IRC calculations, B2GP-PLYP was excluded from the calculations. Additionally, we also performed IRC calculations with M06 and M06-2X. A locally modified version of GAMESS was used to calculate the IRC path of the reaction with BLYP-VBV, PBE-VBV, and TPSS-VBV. All calculations were carried out with the 6-311++G(2df,2p) basis set, without BSSE corrections, as explained in section 2, with the results being shown in Table 6.

The first thing that should be mentioned is the difference between the BH&HLYP results of Table 6 and the ones obtained in ref 46. The explanation is simple, as the BH&HLYP energies in our previous work were calculated as single-point energies for geometries optimized at the

CASSCF(11,11)/6-311++G(2df,2p) level, while in Table 6 all energies concerning a specific functional are the result of geometry optimizations and IRC calculations with the same functional.

The IRC path calculated with the M06-HF functional is the one that resembles most the ab initio CASSCF calculations, since it is the only one that optimizes MIN_2 and SP_3 . However, the relative energies between M06-HF and CASPT2 differ quite dramatically in some points. Additionally, and after many runs, the SP_2 saddle point could not be optimized. In fact, this saddle point could not also be optimized with M06-2X and PBE-VBV, which means that these three functionals generate an extremely flat PES in this region. This problem was also addressed in ref 46. This saddle point connects two isomers with the same energy, MIN_1 ; each of them has the HO_2 fragment tilted to one of the ozone extreme oxygen atoms, MIN_{IL} and MIN_{IR} (see Figure 3 of ref 44 and Figure 1 of ref 46). In the ab initio IRC calculations, MIN_{IL} is connected to SP_1 , while no minimum structure is associated with the IRC path in the direction of the reactants coming from SP_4 . In our KS-DFT calculations, things are slightly different. Except for M06-HF, which has MIN_{IL} associated with SP_1 and SP_4 , the remaining functionals shown in Table 6 have MIN_{IR} associated with SP_1 and MIN_{IL} associated with SP_4 .

Another interesting aspect of the KS-DFT calculations is related to the geometry of the HO_3 fragment in MIN_3 . Ab initio studies have shown that the equilibrium geometry of the hydrogen trioxy radical is strongly dependent on the theoretical method used (see ref 101 and references therein). Three minimum structures are known, one is *gauche*- HO_3 ,¹⁰² with the hydrogen atom out of the plane formed by the three oxygen atoms, and the other two are the planar isomers *cis*- and *trans*- HO_3 .¹⁰³ The HO_3 geometries in MIN_3 obtained with the functionals of Table 6 also show this kind of dependence. BLYP-VBV and PBE-VBV calculate a *cis*- HO_3 structure and M06-HF a *cis*-like- HO_3 structure (one of the atoms is slightly off the plane). M06 and M06-2X calculate a *trans*-like- HO_3 structure, while BH&HLYP and TPSS-VBV originate a *gauche*- HO_3 structure. Table 7 shows the geometric parameters of HO_3 obtained experimentally¹⁰⁴ and calculated at different levels of theory. Besides the already mentioned differences in the dihedral angles ($D(O_1O_2O_3H)$), these defining the *cis*-, *trans*-, or *gauche*-type of structures, the largest differences are observed in the long O_2O_3 bond and in the O_2O_3H angle, consistent with recently reported theoretical work.^{101,105,106}

Looking at the energetics of the stationary points, one can see that the new functionals, while improving the description of SP_1 and SP_4 , still maintain a reasonable quality throughout the remaining structures. This can clearly be seen with BH&HLYP and BLYP-VBV, which differ only by the amount of exact exchange in each of them, 50% and 40.4%, respectively. This decrease of exact exchange has a large effect on SP_1 (decrease of 1.98 kcal mol⁻¹) and SP_4 (decrease of 3.19 kcal mol⁻¹) and a rather small effect on the remaining points, the largest being of 0.59 kcal mol⁻¹ for MIN_3 . In fact, the KS-DFT energies of this minimum are always below the CASPT2 value by at least ~ 3 kcal

Table 7. Comparison between the Geometric Parameters of HO₃ Obtained Experimentally¹⁰⁴ and Theoretically^a

method	<i>d</i> (O ₁ O ₂)	<i>D</i> (O ₂ O ₃)	<i>d</i> (O ₃ H)	<i>A</i> (O ₁ O ₂ O ₃)	<i>A</i> (O ₂ O ₃ H)	<i>D</i> (O ₁ O ₂ O ₃ H)
expt ¹⁰⁴	1.225	1.688	0.972	111.02	90.04	180.00
MRCI ¹⁰⁵	1.233	1.647	0.960	107.40	96.60	180.00
CASSCF(11,11) ⁴⁶	1.259	1.476	0.946	108.77	100.34	-99.91
M06-HF	1.250	1.397	0.969	111.41	101.99	-5.68
M06-2X	1.241	1.452	0.968	109.42	100.08	-154.23
M06	1.228	1.494	0.969	109.69	99.28	-172.46
BH&HLYP	1.251	1.424	0.958	111.04	101.97	-84.62
BLYP-VBV	1.250	1.439	0.966	111.96	101.48	0.00
PBE-VBV	1.239	1.412	0.963	111.91	101.57	0.00
TPSS-VBV	1.246	1.430	0.960	111.08	101.52	-86.14

^a The MRCI geometry was calculated at the MRCI/6-311+G(d,p)//CASSCF(13,13) level¹⁰⁵ (*trans*-HO₃), while the remaining geometries concern the HO₃ fragment of MIN₃ optimized at the CASSCF(11,11)⁴⁶ and KS-DFT level with the 6-311++G(2df,2p) basis set. Distances (*d*, *D*) are in Angstroms, and angles (*A*, *D*) are in degrees.

mol⁻¹, reaching a high error of ~10 kcal mol⁻¹ with M06-HF and M06-2X. Note also that the **SP**₁ and **SP**₄ barrier heights of Table 6 are slightly below the ones of Tables 4 and 5. This is mainly due to the fact that in Table 6 the relative energies were calculated assuming the reactants as a supermolecule, with the fragments separated by 150 Å, as carried out in ref 46, instead of using eq 1. In fact, if these barrier heights are recalculated with the three new functionals by performing single-point energy calculations with the AVTZ basis set, one obtains for **SP**₁ and **SP**₄, respectively, 7.71 and 3.28 kcal mol⁻¹ for BLYP-VBV, 7.49 and 2.50 kcal mol⁻¹ for PBE-VBV, and 8.06 and 4.03 kcal mol⁻¹ for TPSS-VBV, generally improving further the agreement with the CASPT2(11,11)/AVTZ results.

It recently came to our attention a publication concerning the integration grid errors of the M06 suite of functionals,¹⁰⁷ where it is concluded that the grid errors arise from integration errors in the exchange component of the energy. This becomes even more problematic for M06-HF, where the grid errors in predicting reaction energies in a set of 34 organic reactions are shown to go from -6.7 to 3.2 kcal mol⁻¹. Our computations with the GAMESS package were performed with its default grid (96 radial points and 288 angular points (96,288)), and so we performed a separate set of calculations with a larger grid, namely (200,1202), to check if there was any major error associated with our calculations. All absolute energies calculated with the larger grid were shifted positively by less than 1 × 10⁻⁴ E_h (~0.06 kcal mol⁻¹) from the initial energies, a small error that essentially disappears when performing energy differences.

4. Conclusions

In this work we performed a KS-DFT computational benchmark study of the reaction between ozone and the hydroperoxyl radical, comparing the results with the ones resulting from our previous ab initio study.⁴⁶ Because of the nature of our investigation, we used exchange-correlation functionals from all rungs of the “Jacob’s ladder” of density functional approximations in order to assess the performance of a wide range of functionals in describing this particular reaction. Our main concern was to evaluate the quality of the KS-DFT barrier heights of the oxygen- and hydrogen-abstraction mechanisms, since we are primarily interested in the dynamics of this reaction. The best functionals were subsequently used to calculate the remaining stationary points

along the reaction coordinate and to compare the energetics with the CASPT2 calculations.

The barrier heights were shown to be improved using functionals incorporating exact exchange (fourth and fifth rungs of the ladder), as a consequence of the decrease of the SIE in the saddle-point geometries. However, in our first batch of calculations, none of the exchange-correlation functionals lead to good barrier heights for both saddle points, **SP**₁ and **SP**₄. This difficulty is probably extendable in the study of reactions with such complex electronic structure features. We then proceeded by investigating the relation between the fraction of exact exchange in three one-parameter hybrids and the imaginary frequencies of the saddle points. By making the KS-DFT imaginary frequencies match the **SP**₁ ab initio one, we obtained *X* ≈ 40% for the three new functionals (BLYP-VBV, PBE-VBV, and TPSS-VBV), improving both barrier heights. This fraction of exact exchange is very similar to the one existing in functionals that were fitted to barrier heights and are very efficient in thermochemical kinetics.^{98,99} For this reason, it is possible that these three new functionals are also generally suited for kinetics, but to confirm this hypothesis it would be necessary to test them against several databases. However, we stress that our focus is not on the new functionals but rather on the methodology followed to obtain them, as it allowed us to improve our KS-DFT results with a computationally fast and chemically driven procedure. The method also gave us some latitude in which exchange and correlation functionals to choose from when building the final functional. This happened to be quite useful, since our needs were not satisfied by any of the functionals defined in the quantum chemistry packages available in our group.

Finally, we would also like to point out that we recognize the subjectivity inherent to the choice of ab initio method to which the KS-DFT results should be compared to. For this reaction, considering its number of atoms and electronic structure details,⁴⁶ the CASPT2(11,11)/AVTZ//CASSCF(11,11)/6-311++G(2df,2p) level of theory was pretty much the best that we could do with the available computational power at our disposal. This of course does not necessarily mean that those will be, henceforth, the best ab initio calculations available for this reaction. In fact, some studies reveal that MRPT is problematic,⁴⁸⁻⁵⁰ while in other cases it shows a lower accuracy when compared to certain functionals,^{51,52} even for single-reference systems. Some

caution should then be exercised when using MRPT results. For example, a plausible scenario (among many other possibilities) would be one in which the MPW1K functional (or, due to their similarity, PBE-VB) could be the best one to reproduce the barrier heights calculated at some higher level of theory, say multireference configuration interaction calculations. Nevertheless, we think that this study shows a fairly reliable and fast alternative to perform electronic structure calculations for studying the HO₂ + O₃ reaction.

Acknowledgment. The authors acknowledge funding from Fundação para a Ciência e a Tecnologia, Portugal (contracts PTDC/QUI-QUI/099744/2008, PTDC/AAC-AMB/099737/2008, and SFRH/BPD/40807/2007).

References

- Parr, R. G.; Yang, W. *Density-Functional Theory Of Atoms and Molecules*; Oxford University Press: New York, 1989.
- Hohenberg, P.; Kohn, W. *Phys. Rev.* **1964**, *136*, B864.
- Kohn, W.; Sham, L. J. *Phys. Rev.* **1965**, *140*, A1133.
- Scuseria, G. E.; Staroverov, V. N. Progress in the development of exchange-correlation functionals. 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; p 669.
- Perdew, J. P.; Ruzsinszky, A.; Tao, J.; Staroverov, V. N.; Scuseria, G. E.; Csonka, G. I. *J. Chem. Phys.* **2005**, *123*, 062201.
- Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. *J. Phys. Chem.* **1994**, *98*, 11623.
- Perdew, J. P.; Schmidt, K. Jacob's ladder of density functional approximations for the exchange-correlation energy. In *Density Functional Theory and Its Application to Materials*; Van Doren, V., Van Alsenoy, C., Geerlings, P., Eds.; AIP: Melville, NY, 2001; p 1.
- Jones, R. O.; Gunnarsson, O. *Rev. Mod. Phys.* **1989**, *61*, 689.
- Delley, B. *J. Chem. Phys.* **1991**, *94*, 7245.
- Andzelm, J.; Wimmer, E. *J. Chem. Phys.* **1992**, *96*, 1280.
- Johnson, B. G.; Gill, P. M. W.; Pople, J. A. *J. Chem. Phys.* **1993**, *98*, 5612.
- Kohn, W.; Becke, A. D.; Parr, R. G. *J. Phys. Chem.* **1996**, *100*, 12974.
- Becke, A. D. *J. Chem. Phys.* **1992**, *96*, 2155.
- Becke, A. D. *J. Chem. Phys.* **1992**, *97*, 9173.
- Tao, J.; Perdew, J. P.; Staroverov, V. N.; Scuseria, G. E. *Phys. Rev. Lett.* **2003**, *91*, 146401.
- Harris, J.; Jones, R. O. *J. Phys. F* **1974**, *4*, 1170.
- Gunnarsson, O.; Lundqvist, B. I. *Phys. Rev. B* **1976**, *13*, 4274.
- Langreth, D. C.; Perdew, J. P. *Phys. Rev. B* **1977**, *15*, 2884.
- Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 1372.
- Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648.
- Perdew, J. P.; Ernzerhof, M.; Burke, K. *J. Chem. Phys.* **1996**, *105*, 9982.
- Gritsenko, O. V.; Van Leeuwen, R.; Baerends, E. J. *Int. J. Quantum Chem.* **1996**, *60*, 1375.
- Ernzerhof, M.; Perdew, J. P.; Burke, K. *Int. J. Quantum Chem.* **1997**, *64*, 285.
- Jaramillo, J.; Scuseria, G. E.; Ernzerhof, M. *J. Chem. Phys.* **2003**, *118*, 1068.
- Mori-Sánchez, P.; Cohen, A. J.; Yang, W. *J. Chem. Phys.* **2006**, *124*, 091102.
- Arbuznikov, A. V.; Kaupp, M. *Chem. Phys. Lett.* **2007**, *440*, 160.
- Perdew, J. P.; Ruzsinszky, A.; Csonka, G. I.; Vydrov, O. A.; Scuseria, G. E.; Staroverov, V. N.; Tao, J. *Phys. Rev. A* **2007**, *76*, 040501.
- Perdew, J. P.; Staroverov, V. N.; Tao, J.; Scuseria, G. E. *Phys. Rev. A* **2008**, *78*, 052513.
- Haunschuld, R.; Scuseria, G. E. *J. Chem. Phys.* **2010**, *132*, 224106.
- Perdew, J. P.; Ernzerhof, M. Driving out the self-interaction error. In *Electronic Density Functional Theory: Recent Progress and New Directions*; Dobson, J. F., Vignale, G., Das, M. P., Eds.; Plenum: New York, 1998; p 31.
- Lynch, B. J.; Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2003**, *107*, 1384.
- Zhao, Y.; Pu, J.; Lynch, B. J.; Truhlar, D. G. *Phys. Chem. Chem. Phys.* **2004**, *6*, 673.
- Levine, I. N. Ab initio and density-functional treatments of molecules. In *Quantum Chemistry*, 5th ed.; Prentice-Hall: New Jersey, 2000; p 581.
- Patchkovskii, S.; Ziegler, T. *J. Chem. Phys.* **2002**, *116*, 7806.
- Perdew, J. P.; Zunger, A. *Phys. Rev. B* **1981**, *23*, 5048.
- Zhao, Y.; Lynch, B. J.; Truhlar, D. G. *J. Phys. Chem. A* **2004**, *108*, 4786.
- Duewer, W. H.; Wuebbles, D. J.; Ellsaesser, H. W.; Chang, J. S. *J. Geophys. Res.* **1977**, *82*, 935.
- Crutzen, P. J.; Howard, C. J. *Pure Appl. Geophys.* **1978**, *116*, 497.
- Whitten, R. C.; Borucki, W. J.; Capone, L. A.; Turco, R. P. *Nature* **1978**, *275*, 523.
- Turco, R. P.; Whitten, R. C.; Poppoff, I. G.; Capone, L. A. *Nature* **1978**, *276*, 805.
- Wennberg, P. O.; et al. *Science* **1994**, *266*, 398.
- Monks, P. S. *Chem. Soc. Rev.* **2005**, *34*, 376.
- Varandas, A. J. C.; Zhang, L. *Chem. Phys. Lett.* **2004**, *385*, 409.
- Mansergas, A.; Anglada, J. M. *J. Phys. Chem. A* **2007**, *111*, 976.
- Xu, Z. F.; Lin, M. C. *Chem. Phys. Lett.* **2007**, *440*, 12.
- Viegas, L. P.; Varandas, A. J. C. *J. Chem. Theory Comput.* **2010**, *6*, 412.
- Perdew, J. P.; Ruzsinszky, A.; Constantin, L. A.; Sun, J.; Csonka, G. I. *J. Chem. Theory Comput.* **2009**, *5*, 902.
- Rode, M. F.; Werner, H.-J. *Theor. Chem. Acc.* **2005**, *114*, 309.
- Cramer, C. J.; Włoch, M.; Piecuch, P.; Puzzarini, C.; Gagliardi, L. *J. Phys. Chem. A* **2006**, *110*, 1991.
- Cramer, C. J.; Kinal, A.; Włoch, M.; Piecuch, P.; Gagliardi, L. *J. Phys. Chem. A* **2006**, *110*, 11557.

- (51) Tishchenko, O.; Zhen, J.; Truhlar, D. G. **2008**, *4*, 1208.
- (52) Zhen, J.; Zhao, Y.; Truhlar, D. G. **2009**, *5*, 808.
- (53) Zhao, Y.; Tishchenko, O.; Gour, J. R.; Li, W.; Lutz, J. J.; Piecuch, P.; Truhlar, D. G. *J. Phys. Chem. A* **2009**, *113*, 5786.
- (54) Schmidt, M. W.; Baldrige, K. K.; Boats, J. A.; Elbert, S. T.; Gorgon, M. S.; Jensen, J. H.; Koseki, S.; Matsunaga, N.; Nguyen, K. A.; Su, S.; Windus, T. L.; Dupuis, M.; Montgomery, J., Jr. *J. Comput. Chem.* **1993**, *14*, 1347.
- (55) Neese, F. *ORCA - an ab initio, Density Functional and Semiempirical program package*, Version 2.6-35; University of Bonn, 2008.
- (56) Bode, B. M.; Gordon, M. S. *J. Mol. Graphics Modell.* **1998**, *16*, 133.
- (57) Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098.
- (58) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785.
- (59) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865.
- (60) Adamo, C.; Barone, V. *Chem. Phys. Lett.* **1998**, *298*, 113.
- (61) Adamo, C.; Barone, V. *J. Chem. Phys.* **1999**, *110*, 6158.
- (62) Zhao, Y.; Truhlar, D. G. *Theor. Chem. Acc.* **2008**, *120*, 215.
- (63) Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2008**, *110*, 13126.
- (64) Grimme, S. *J. Chem. Phys.* **2006**, *124*, 034108.
- (65) Karton, A.; Tarnopolsky, A.; Lamere, J. F.; Schatz, G. C.; Martin, J. M. L. *J. Phys. Chem. A* **2008**, *112*, 12868.
- (66) Grimme, S. *J. Comput. Chem.* **2004**, *25*, 1463.
- (67) Grimme, S. *J. Comput. Chem.* **2006**, *27*, 1787.
- (68) Schwabe, T.; Grimme, S. *Acc. Chem. Res.* **2008**, *41*, 569.
- (69) Korth, M.; Grimme, S. *J. Chem. Theory Comput.* **2009**, *5*, 993.
- (70) Gruzman, D.; Karton, A.; Martin, J. M. L. *J. Phys. Chem. A* **2009**, *113*, 11974.
- (71) Snook, I. K.; Per, M. C.; Seyed-Razavi, A.; Russo, S. P. *Chem. Phys. Lett.* **2009**, *480*, 327.
- (72) Steinmann, S. N.; Csonka, G.; Corminboeuf, C. *J. Chem. Theory Comput.* **2009**, *5*, 2950.
- (73) Zhao, Y.; Ng, H. T.; Hanson, E. *J. Chem. Theory Comput.* **2009**, *5*, 2726.
- (74) Flener-Lovitt, C.; Woon, D. E.; Dunning, T. H., Jr.; Girolami, G. S. *J. Phys. Chem. A* **2010**, *114*, 1843.
- (75) Marom, N.; Tkatchenko, A.; Scheffler, M.; Kronik, L. *J. Chem. Theory Comput.* **2010**, *6*, 81.
- (76) Shamov, G. A.; Budzelaar, P. H. M.; Schreckenbach, G. *J. Chem. Theory Comput.* **2010**, *6*, 477.
- (77) Boys, S. F.; Bernardi, F. *Mol. Phys.* **1970**, *19*, 553.
- (78) Xantheas, S. S. *J. Chem. Phys.* **1996**, *104*, 8821.
- (79) Szalewicz, K.; Jeziorski, B. *J. Chem. Phys.* **1998**, *109*, 1198.
- (80) Varandas, A. J. C. *Theor. Chem. Acc.* **2008**, *119*, 511.
- (81) Varandas, A. J. C. DOI:10.1021/jp908835v.
- (82) Johnson, B. G.; Gonzales, C. A.; Will, P. M. W.; Pople, J. A. *Chem. Phys. Lett.* **1994**, *221*, 100.
- (83) Porezag, D.; Pederson, M. R. *J. Chem. Phys.* **1995**, *102*, 9345.
- (84) Csonka, G. I.; Johnson, B. G. *Theor. Chem. Acc.* **1998**, *99*, 158.
- (85) Galano, A.; Alvarez-Idaboy, J. R.; Montero, L. A.; Vivier-Bunge, A. *J. Comput. Chem.* **2001**, *22*, 1138.
- (86) Plesničar, B.; Tuttle, T.; Cerkovnik, J.; Koller, J.; Cremer, D. *J. Am. Chem. Soc.* **2003**, *125*, 11553.
- (87) Wu, A.; Cremer, D.; Plesničar, B. *J. Am. Chem. Soc.* **2003**, *125*, 9395.
- (88) Gräfenstein, J.; Kraka, E.; Cremer, D. *Phys. Chem. Chem. Phys.* **2004**, *6*, 1096.
- (89) Tarnopolsky, A.; Karton, A.; Sertchook, R.; Vuzman, D.; Martin, J. M. L. *J. Phys. Chem. A* **2008**, *112*, 3.
- (90) Lynch, B. J.; Truhlar, D. G. *J. Phys. Chem. A* **2001**, *105*, 2936.
- (91) Zhao, Y.; González-García, N.; Truhlar, D. G. *J. Phys. Chem. A* **2005**, *109*, 2012.
- (92) Su, J. T.; Xu, X.; Goddard III, W. A. *J. Phys. Chem. A* **2004**, *108*, 10518.
- (93) Halkier, A.; Klopper, W.; Helgaker, T.; Jørgensen, P.; Taylor, P. R. *J. Chem. Phys.* **1999**, *111*, 9157.
- (94) Lin, R. J.; Wu, C. C.; Jang, S.; Li, F.-Y. *J. Mol. Model.* **2010**, *16*, 175.
- (95) Pu, J.; Truhlar, D. G. *J. Chem. Phys.* **2002**, *116*, 1468.
- (96) Jeanvoine, Y.; Spezia, R. *J. Mol. Struct.: THEOCHEM* **2010**, *954*, 7.
- (97) Staroverov, V. N.; Scuseria, G. E.; Tao, J.; Perdew, J. P. *J. Chem. Phys.* **2003**, *119*, 12129.
- (98) Lynch, B. J.; Fast, P. L.; Harris, M.; Truhlar, D. G. *J. Phys. Chem. A* **2000**, *104*, 4811.
- (99) Zhao, Y.; Lynch, B. J.; Truhlar, D. G. *J. Phys. Chem. A* **2004**, *108*, 2715.
- (100) Adamo, C.; Barone, V. *J. Chem. Phys.* **1998**, *108*, 664.
- (101) Braams, B. J.; Yu, H.-G. *Phys. Chem. Chem. Phys.* **2008**, *10*, 3150.
- (102) Dupuis, M.; Fitzgerald, G.; Hammond, B.; Lester, W. A., Jr.; Schaefer III, H. F. *J. Chem. Phys.* **1986**, *84*, 2691.
- (103) Jungkamp, T. P. W.; Steinfeld, J. H. *Chem. Phys. Lett.* **1996**, *257*, 15.
- (104) Suma, K.; Sumiyoshi, Y.; Endo, Y. *Science* **2005**, *308*, 1885.
- (105) Mansergas, A.; Anglada, J. M.; Olivella, S.; Ruiz-López, M. F.; Martins-Costa, M. *Phys. Chem. Chem. Phys.* **2007**, *9*, 5865.
- (106) Varner, M. E.; Harding, M. E.; Gauss, J.; Stanton, J. F. *Chem. Phys. Lett.* **2008**, *346*, 53.
- (107) Wheeler, S. E.; Houk, K. N. *J. Chem. Theory Comput.* **2010**, *6*, 395.

JCTC

Journal of Chemical Theory and Computation

E/Z Energetics for Molecular Modeling and Design

John P. Terhorst and William L. Jorgensen*

Department of Chemistry, Yale University, New Haven, Connecticut 06520-8107

Received July 19, 2010

Abstract: Thermochemical data have been obtained from G3B3 quantum mechanical calculations for 18 prototypical organic molecules, which exhibit *E/Z* conformational equilibria. The results are fundamentally important for molecular design including evaluation of structures from protein–ligand docking. For the 18 *E/Z* pairs, relative energies, enthalpies, free energies, and dipole moments are reported; the *E* – *Z* free-energy differences at 298 K range from +8.2 kcal/mol for 1,3-dimethyl carbamate to –6.4 kcal/mol for acetone oxime. A combination of steric and electronic effects can rationalize the variations. Free energies of hydration were also estimated using the GB/SA continuum solvent model. These results indicate that differential hydration is unlikely to qualitatively change the preferred direction of the *E/Z* equilibria, though further study with free-energy methods using explicit solvent is desirable.

Introduction

Knowledge of the conformational energetics of small molecules is essential in many areas of chemistry including organic synthesis and molecular design.¹ The conformational preferences for small molecules are well known to carry over to macromolecular structures, e.g., the ca. 3 kcal/mol preference for the *Z* conformer of *N*-methylacetamide relative to the *E* alternative is primarily responsible for the rarity of *cis*-peptide bonds in proteins.² The present study focuses on such molecules where rotation about a single bond leads to *E* and *Z* conformers that are energetically well separated by an intervening potential-energy barrier. Besides amides, molecules in this category include other derivatives of carboxylic acids, aldehydes, or ketones such as esters, carbamates, carbonates, ureas, amidines, hydrazones, and oximes. The importance of these functional groups is enhanced by their common occurrence in combinatorial libraries, commercial screening collections, and molecules of pharmacological interest.

Furthermore, in seeking enzyme inhibitors through de novo design or virtual screening,^{3,4} questions often arise about the likelihood of *E* and *Z* conformers. For example, in docking studies, one is regularly confronted with computed structures for complexes, ‘poses’, such as in Figure 1a, where the ligand features an *E* or *Z* conformation. The scoring with docking software is still improving and such poses may score well,⁵

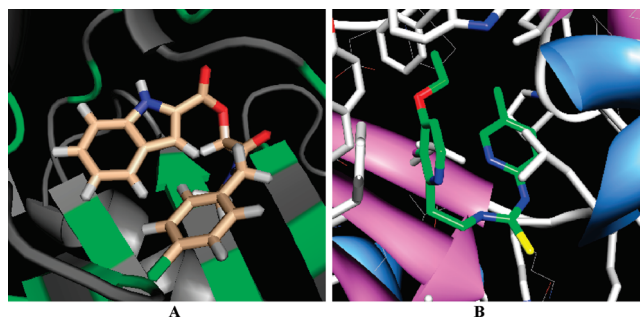


Figure 1. (A) Structure of an ester-containing molecule docked into HIV-1 reverse transcriptase (RT), and (B) the 1dt crystal structure of an analog of trovirdine bound to HIV-RT illustrating an *E,Z* conformer for a thiourea moiety.

though the *E* conformer for the ester in this case is unreasonable.¹ Alternately, one may be confronted with a crystal structure, such as in Figure 1b, where the thiourea moiety is in an *E,Z* configuration.⁶ If one thought that there was an associated energetic penalty, alternative designs might be pursued to achieve enhanced potency. Given many such examples, we pursued energetic clarification through reliable quantum mechanical calculations on prototypical molecules featuring *E* and *Z* conformers. The findings are also valuable as a basis for the improvement of scoring functions for docking software,⁷ refinement of crystal structures, and development of molecular mechanics force fields for use in modeling organic and biomolecular systems.^{8,9}

* Corresponding author e-mail: william.jorgensen@yale.edu.

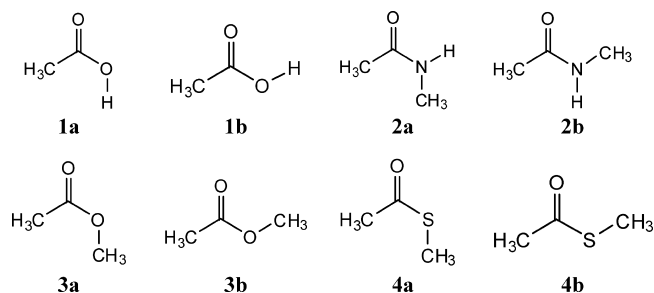


Figure 2. Molecules in the RCOX set.

Though there have been prior computational studies of molecules featuring *E/Z* equilibria, most studies have focused on one or two functional groups using Hartree–Fock (HF), B3LYP-based density functional, or second-order Møller–Plesset (MP2) theory.^{10–30} Some classic studies include those of Wiberg and co-workers on formic acid, acetic acid, methyl formate, and methyl acetate.¹⁰ *N*-Methylacetamide has also received much attention owing to its status as a model for the peptide bond.^{11,20,23,24,30} The need for quantum mechanical investigations in this area is enhanced by the fact that experimental studies of *E/Z* equilibria are often challenging owing to a very small population of the higher energy conformer. For example, the first experimental observation of the *E* conformer of acetic acid was not made until 2003.³¹ Thus, for broader coverage of *E/Z* equilibria at a higher and consistent level of theory, the 18 pairs of conformers illustrated in Figures 2–4 have been examined here using composite ab initio methods.

Computational Details

All ab initio and DFT calculations were carried out using the Gaussian03 program.³² The G3 and G3B3 methods were applied to compute structures, dipole moments, vibrational frequencies, energies at 0 K and enthalpies and free energies at 298 K.^{33,34} With the G3 method, the initial geometry optimization and vibrational frequency and zero-point energy calculations are performed at the 6-31G(d) level. The geometry is then refined including electron correlation at the MP2(full)/6-31G(d) level. A series of single-point energy calculations follows, using MP2/G3large (a basis set with core correlation), MP4/6-31G(d), and QCISD(T)/6-31G(d),

with spin–orbit and other higher corrections. The G3B3 approach particularly improves the initial geometry, vibrational frequencies, and zero-point energy by starting with a B3LYP/6-31G(d) geometry optimization. The increase in computer time for G3B3 over G3 for molecules of the present size is usually less than 50%.

Estimates of free energies of hydration were made for all conformers using the generalized Born/surface area approach, as implemented in the BOSS program.^{35,36} Structures were optimized using the OPLS/CM1A force field,⁹ and the GB/SA calculations were performed with CM1A atomic charges scaled by 1.07.³⁶

Results and Discussion

***E/Z* Conformers.** The 18 pairs of conformers that were investigated are shown in Figures 2–4. The RCOX set consists of a prototypical carboxylic acid, secondary amide, ester, and thioester. The RXCOYR set contains a urea, thiourea, carbamate (urethane), and carbonate, while the C=C&N set covers an enamine, an enol ether, amidines, hydrazones, and oximes. Conformer **a** is *E* and conformer **b** *Z* for each pair. For amine derivatives, secondary cases RNHCH₃ have been considered; *E* and *Z* are also well defined for tertiary cases RNR'R'', but the *E/Z* preferences for them are generally well predicted by steric considerations. It should also be noted that conformers **7b** and **8b** are the same, which simplifies the presentation of results. For the RCOX set, both G3 and G3B3 calculations were performed, while the RXCOYR and C=C&N sets were investigated only using G3B3.

Results for the RCOX Set. For 1–4, the G3 and G3B3 results are shown in Table 1. In all cases, the relative values are given for conformer **a** minus conformer **b** (*E* – *Z*). The G3 and G3B3 energetic results generally agree to within 0.1 kcal/mol. The thermal corrections to the vibrational energy are also almost the same for both conformers, so there is little difference between the results for ΔE (0 K) and ΔH (298 K). The computed entropy changes are generally small, though there can be some sensitivity to the treatment of low-frequency vibrations.

For acetic acid (**1**), the *E* conformer is found to be 5.1 kcal/mol higher in energy than the *Z* form from the G3 and

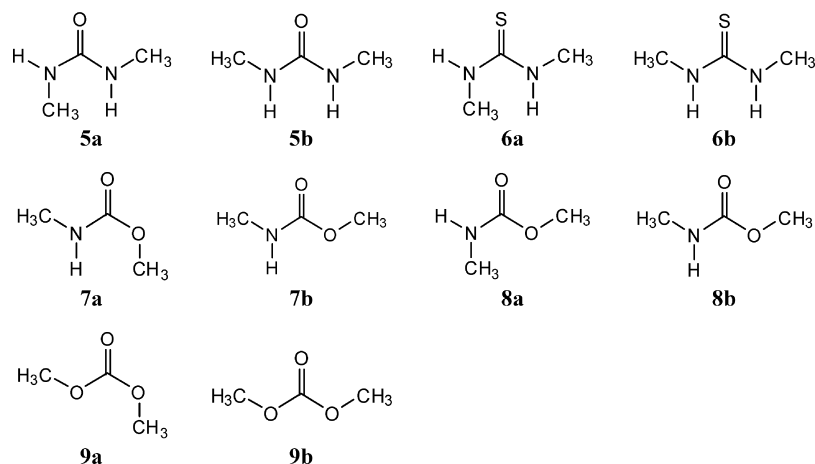


Figure 3. Molecules in the RXCOYR set.

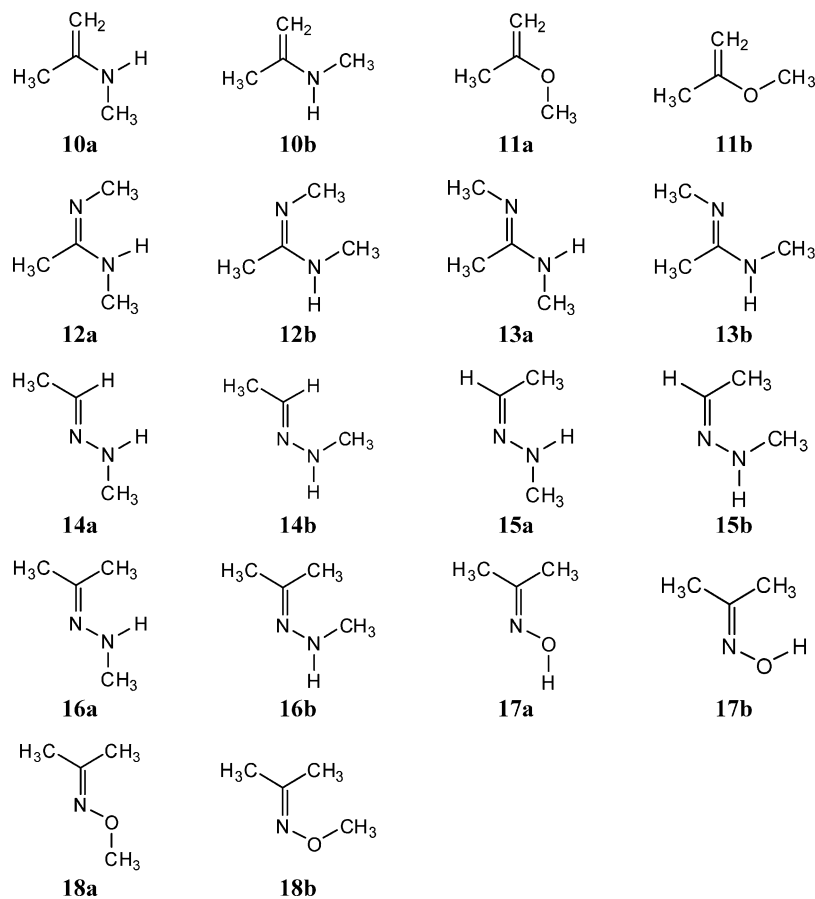


Figure 4. Molecules in the C=C&N set.

Table 1. Computed Differences in Energies (kcal/mol) and Dipole Moments (D) from G3 and G3B3 Calculations for the RCOX Set

pair	ΔE (0 K)	ΔH (298 K)	ΔG (298 K)	$\Delta\mu$
G3				
1	5.08	5.11	5.15	2.93
2	2.42	2.22	3.11	0.32
3	7.48	7.46	7.47	3.10
4	4.63	4.43	4.60	3.14
G3B3				
1	5.11	5.11	5.27	2.92
2	2.34	2.26	2.67	0.32
3	7.42	7.41	7.42	3.10
4	4.63	4.41	4.38	3.19

G3B3 calculations. This is in accord with an MP4/cc-pVTZ result of 5.38 kcal/mol,²⁵ while lower levels of ab initio theory generally give larger differences.¹⁰ An experimental result is not available for comparison, though the *E* conformer has been detected in an argon matrix at 8 K.³¹ The best estimate for the energy difference for formic acid is about 1 kcal/mol smaller at 4.21 kcal/mol.²² For *N*-methylacetamide (**2**), the present results concur with other high-level calculations and experiments that the enthalpy difference at 298 K is in the 2.1–2.5 kcal/mol range.^{11,20,23,24,30} The difference diminishes to 1.0–1.2 kcal/mol for *N*-methylformamide owing to reduced steric crowding in the *E* form.^{20,37}

Similarly, the G3 and G3B3 results for methyl acetate are in-line with the energy difference of 7.72 kcal/mol from

LMP2/cc-pVTZ(-f) calculations,²⁰ while again older values are somewhat higher.^{10,15,16} For methyl formate, the LMP2 energy difference is reduced to 5.35 kcal/mol.²⁰ Besides the steric effects favoring *Z*, the *E* conformer of carboxylic acids and esters is also destabilized by unfavorable dipole–dipole interactions or lone-pair–lone-pair repulsion between the oxygen atoms.^{10,11} As indicated in Table 1, the dipole moments for *E* acids and esters are ca. 3 D higher than for the *Z* forms. Overall, the population of *E* carboxylic esters is generally vanishingly low and drawing or invoking acyclic esters in this conformation is improper.³⁸ The *E*–*Z* energy difference for the corresponding thioester (**4**) moderates to 4.6 kcal/mol owing in part to the longer C–S than C–O bonds, which diminishes the 1,4-CC steric penalty for the *E* conformer. Again, the difference is expected to be less for methyl thioformate, which is confirmed by NMR studies indicating a free-energy difference of ca. 1.3 kcal/mol.³⁷

Results for the RXCOYR and C=C&N Sets. The G3B3 results for the remaining pairs are given in Table 2. The results are largely understandable in terms of the strong preference for the *Z* conformers for **1–3** and additional steric and electronic effects, as presented below.

Interestingly, in comparison to *N*-methylacetamide, the *Z,Z* over *E,Z* energetic preference for 1,3-dimethylurea (**5**) diminishes to 1.06 kcal/mol, and the *E,Z* conformer of 1,3-dimethylthiourea (**6a**) is actually favored by 0.17 kcal/mol. In prior work, MP2/aug-cc-pVDZ results favored the *Z* conformer of methylurea and methylthiourea by 1.25 and 0.70 kcal/mol,^{28,29} and MP2/6-31G(d) results preferred *Z,Z*

Table 2. Computed Differences in Energies (kcal/mol) and Dipole Moments (D) from G3B3 Calculations for the RXCOYR and C=C&N Sets

pair	ΔE (0 K)	ΔH (298 K)	ΔG (298 K)	$\Delta\mu$
RXCOYR				
5	1.06	1.03	1.09	0.50
6	-0.17	-0.09	-0.55	0.80
7	7.47	7.30	8.18	3.05
8	1.24	1.15	1.75	0.34
9	3.03	2.99	3.09	3.60
C=C&N				
10	2.67	2.65	2.60	-0.13
11	4.47	4.61	4.01	1.27
12	-4.06	-4.05	-3.95	0.31
13	3.13	3.00	3.44	-0.04
14	-0.16	-0.03	-0.35	-0.01
15	-3.30	-3.21	-3.28	0.37
16	-2.54	-2.43	-2.69	0.54
17^a	-6.04	-6.00	-6.35	-2.94
18^a	-19.07	-18.34	-20.40	-2.66

^a The planar Z form **b** is a transition state.

over *E,Z* for 1,3-dimethylurea by 1.72 kcal/mol.¹⁹ The present result for **5** is expected to be more accurate and indicates that there would only be a small intrinsic penalty for incorporating an *E,Z*-urea substructure in a molecular design. Moreover, an *E,Z*-thiourea fragment as in Figure 1B is preferred over the *Z,Z* alternative. In fact, there have been extensive NMR studies of the conformational equilibria for **6** in multiple solvents with the conclusion that the *E,Z* conformer is lower by ca. 1 kcal/mol in free energy than the *Z,Z* conformer and that the *E,E* form is not populated.²¹ The electronic energy from MP2/cc-pVDZ calculations without zero-point or other corrections in that study appears to lead to the wrong qualitative conclusion by favoring the *Z,Z* conformer by 0.38 kcal/mol.²¹ In summary, the *E,Z* conformer for ureas is relatively more favorable than the *E* conformer of secondary amides, and the *E,Z* conformer for the prototypical 1,3-dialkylthiourea **6** is the lowest in energy. In view of the small differences in dipole moments for **5** and **6** in Table 2, the preferences are expected to not be strongly influenced by medium effects. A possible contributor to the increased favorability of the *E,Z* geometry in the ureas is π -electron donation (amide resonance $^+N=C-X^-$), which increases the partial negative charge on the oxygen or sulfur atom and improves the electrostatic interaction with the *syn*-hydrogen on nitrogen in the *E* substructure.

The results for **7–9** in Table 2 present an interesting contrast. For 1,3-dimethyl carbamate **7**, rotation of the methoxy group to the *E* form is similarly unfavorable as for the ester **3**, while rotation of the *N*-methyl group in going from **8b** to **8a** is about 1 kcal/mol less costly than for the amide **2**. The relative G3B3 energies for the three conformers of the carbamate *Z,Z* (**7b**), *Z,E* (**7a**), and *E,Z* (**8a**) are 0.0, 7.47, and 1.25 kcal/mol, respectively. Thus, as for the urea **5**, the penalty for rotation of the *N*-methyl group in the carbamate to the *E* form is not large; however, an *E* geometry for the ester fragment remains too high in energy for significant population under normal conditions. The possibility for an *E*-ester substructure is significantly improved for dimethyl carbonate (**9**), for which the *E,Z* conformer is only 3.03 kcal/mol higher in energy than the *Z,Z* form. The

Table 3. G3B3 Results for Key Dihedral Angles (deg)

conformer	angle	φ	conformer	φ
1a	CCOH	0.0	1b	180.0
2a	CCNC	9.8	2b	179.9
3a	CCOC	0.3	3b	179.9
4a	CCSC	0.0	4b	178.3
5a	NCNC	20.4	5b	169.4
6a	NCNC	7.0	6b	173.9
7a	NCOC	5.6	7b	179.9
8a	OCNC	9.8	8b	179.9
9a	OCOC	0.0	9b	180.0
10a	CCNC	40.8	10b	170.6
11a	CCOC	37.1	11b	180.0
12a	NCNC	165.0	12b	5.0
13a	NCNC	148.5	13b	9.0
14a	CNNC	152.0	14b	22.3
15a	CNNC	159.7	15b	69.2
16a	CNNC	161.3	16b	79.7
17a	CNOH	179.9	17b^a	0.0
18a	CNOC	180.0	18b^a	1.4

^a Transition state.

4–5 kcal/mol diminution relative to **3** or **7** likely stems from destabilization of the *Z,Z* conformer by repulsion between the lone pairs on the methoxy oxygen atoms. Previous MP2/6-31G(d) results for the relative electronic energies of the *Z,Z*, *E,Z*, and *E,E* conformers of **9** are 0.0, 3.36, and 26.73 kcal/mol.¹⁸

Turning to the molecules in Figure 4, **10** (*N*-methyl-2-aminopropene) and **11** (2-methoxypropene) are the olefinic analogs of **2** and **3**. The energetic preference remains the same, significantly favoring the *Z* conformers by 2.67 (**10**) and 4.47 kcal/mol (**11**). Thus, the 1,4-CH₃/CH₃ interaction appears to continue to dominate, while the larger energy difference for the ester **3** than the enol ether **11** can be attributed to the addition of the lone-pair repulsion between the oxygens for the *E* conformer of the ester (**3a**). On the basis of the results mentioned above for formic acid vs acetic acid derivatives, the *E* – *Z* energy differences for the corresponding vinyl analogs of **10** and **11** should be reduced by ca. 1.0 and 2–3 kcal/mol, respectively. Indeed, MP3/6-31G results provide an *E* – *Z* energy difference of about 2.0 kcal/mol for methyl vinyl ether,³⁹ and we find 1.74 kcal/mol for $\Delta E(0\text{ K})$ using G3B3. It should be noted that the *E* conformers for **10** and **11** are not planar; the G3B3 results for the H₃C–C–X–CH₃ dihedral angles are 40.8° and 37.1° for **10a** and **11a**. Thus, these conformers may be described as *skew*. Amine nitrogens are also somewhat paramidialized in all structures, so, for example, the H₃C–C–N–CH₃ dihedral angle in **10b** is 170.6°; however, the H₃C–C–O–CH₃ dihedral angle in **11b** is 180°. The results for the corresponding dihedral angles for all conformers are listed in Table 3.

For **12** and **13** in Figure 4, the structures represent the four conformers for *N,N'*-dimethylacetamide. The *trans*-(*Z*) conformer **13b** can be argued to be the most analogous to (*Z*)-*N*-methylacetamide (**2b**) and is the lowest in energy. The relative energies, $\Delta E(0\text{ K})$, for the other conformers are 1.07, 3.13, and 5.13 kcal/mol for **12a**, **13a**, and **12b**, respectively, at the G3B3 level. The *E* – *Z* energy difference in Table 2 for the **13** pair is also just a little greater than for **2**, possibly reflecting diminished electrostatic attraction for

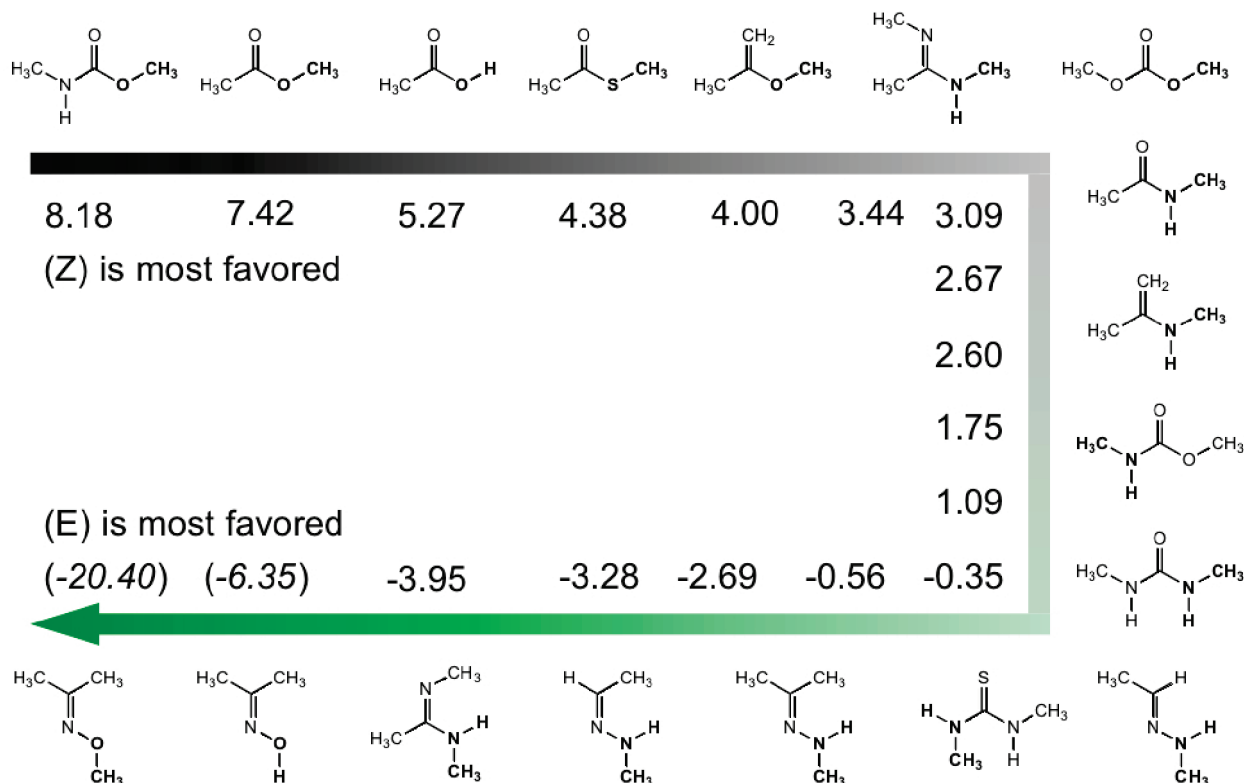


Figure 5. Summary of the G3B3 *E/Z* free-energy differences (kcal/mol). The preferred conformation is shown, and the fragment that is rotated is highlighted in bold.

$N\cdots HN$ in **13a** than for $O\cdots HN$ in **2a**. The 1,5- CH_3-CH_3 interaction in **12b** is particularly destabilizing as it is similar to a *syn*-pentane interaction, so this conformer is not competitive. Overall, two low-energy conformers are apparent for the dimethylamide, **13b** and **12a**.

Similarly, for **14** and **15** in Figure 4, the four structures are the conformers for the *N*-methyl hydrazone of acetaldehyde, while **16a** and **16b** are the *E* and *Z* possibilities for the *N*-methyl hydrazone of acetone. For **14** and **15**, the lowest energy conformer is **14a** and the relative energies, $\Delta E(0\text{ K})$, are 0.16, 0.34, and 3.63 for **14b**, **15a**, and **15b**. Thus, the first three conformers are very close in energy with only **15b** being uncompetitive owing to the 1,5- CH_3-CH_3 interaction. It is also then easy to predict that *Z* conformer **16b** is higher in energy than the *E* alternative **16a**; the difference of 2.54 kcal/mol is a little smaller in magnitude than for the **15** pair. A message from this for molecular design is that the conformational diversity of hydrazones of ketones is much less than for hydrazones of aldehydes.

Finally, the oxime **17** and *O*-methyl oxime **18** of acetone were considered. In both cases, the planar *Z* form was found to be a transition state with the G3B3 calculations and only the *E* structures are energy minima. The *Z* transition states are 6.04 (**17**) and 19.07 (**18**) kcal/mol higher in energy than the *E* conformers. The *Z* structures are destabilized by electrostatic repulsion between the lone-pair electrons on N and O and by the *syn*-pentane-like interaction in **18b**. The status of the *Z* structure for acetoxime **17** is sensitive to the computational level. For example, we find with B3LYP/6-31G(d) energy minimizations and vibrational frequency calculations that the $C=N-O-H$ planar *Z* structure is a

shallow energy minimum; it is 6.16 kcal/mol above the *E* conformer and is separated from conversion to the *E* form by a barrier of 2.0 kcal/mol at a dihedral angle near 70° . The hydroxyl hydrogen is constrained between two of the hydrogens on the *syn*-methyl group, which stagger the $C=N$ bond. However, the G3B3 results state that for both the oxime **17** and *O*-methyl oxime **18**, only the *E* conformational energy well exists. Besides the common occurrence of oximes in screening collections, interest in them also continues as the substrates for Beckmann rearrangements. In this case, the dominance of the *E* conformers is relevant for proposed mechanistic schemes.⁴⁰

Summary of *E/Z* Results. A summary of the *E/Z* free-energy differences for all conformer pairs is provided in Figure 5. A positive difference indicates that the *Z* conformer is favored, and a negative difference indicates that the *E* conformer is favored. Some general rules are evident. (1) For rotation about $C-X$ ($X = OR, SR, NHR$) single bonds in $O=C-X$, $N=C-X$, and $C=C-X$ substructures, *Z* conformers are normally preferred in the absence of significant steric effects that preferentially destabilize the *Z* conformer, especially *syn*-pentane-like interactions. (2) The *Z* preference diminishes in the order $X = OR > OH > SR > NHR$. The *Z* preference is also diminished for thione derivatives, $S=C-X$, and through additional conjugation as in ureas. (3) Rotation about the $N-N$ and $N-O$ bonds in hydrazones and oximes favors the *E* conformers, especially when reinforced by a *syn*-pentane-like interaction in the *Z* form.

Concerning dipole moments, there is a general correlation in Tables 1 and 2 such that the conformer with the larger dipole moment is normally higher in energy than the one

Table 4. Computed E - Z Free-Energy Differences (kcal/mol) in the Gas Phase and in Aqueous Solution at 298 K^a

pair	ΔG_{gas}	$\Delta\mu$	$\Delta\Delta G_{\text{hyd}}$	ΔG_{aq}
7	8.18	3.05	-0.52	7.66
3	7.42	3.10	-0.60	6.82
1	5.27	2.92	-2.39	2.88
4	4.38	3.19	-0.12	4.26
11	4.01	1.27	-1.82	2.18
13	3.44	-0.04	1.75	5.19
9	3.09	3.60	0.26	3.35
2	2.67	0.32	1.43	4.10
10	2.60	-0.13	0.29	2.89
8	1.75	0.34	0.37	2.12
5	1.09	0.50	0.00	1.09
14	-0.35	-0.01	-1.02	-1.37
6	-0.55	0.80	1.28	0.73
16	-2.69	0.54	-1.06	-3.75
15	-3.28	0.37	-0.36	-3.64
12	-3.95	0.31	1.79	-2.16
(17)	-6.35	-2.94	-0.11	-6.46
(18)	-20.40	-2.66	-1.40	-21.80

^a G3B3 results in the gas phase; hydration effect ($\Delta\Delta G_{\text{hyd}}$) from GB/SA calculations. Ordered by decreasing ΔG_{gas} .

with the smaller dipole moment. This is reasonable based on electrostatic considerations and contributes to the general preference for Z conformers. The largest differences in dipole moments ($\Delta\mu = \mu_E - \mu_Z$) are 3–4 D, and these correspond to cases where the E conformer is higher in energy by 3–8 kcal/mol. For the oximes, $\Delta\mu$ is ca. -3 D and consistently the E conformers are significantly favored. Of course, steric effects modulate the results such that, for example, $\Delta\mu$ is small for **12**, **15**, and **16**, but the E conformers are strongly favored owing to the 1,5-CH₃-CH₃ interactions in the Z conformers.

GB/SA Results. The gas-phase results for the E/Z preferences can be shifted in different molecular environments, both relatively homogeneous as for a pure solvent and inhomogeneous as in a protein binding site. To gain some sense of magnitude for the former case, free energies of hydration were calculated for all conformers using the OPLS/CM1A force field and GB/SA continuum solvent model.^{9,36} The gas-phase G3B3 results, the GB/SA shifts $\Delta\Delta G_{\text{hyd}}$, and the net ΔG_{aq} for the $Z \rightleftharpoons E$ equilibria in aqueous solution at 298 K are summarized in Table 4. A negative $\Delta\Delta G_{\text{hyd}}$ indicates that the E conformer is predicted to be better hydrated. On the basis of calculations for 399 neutral organic molecules, the average absolute error for free energies of hydration from the GB/SA calculations is expected to be 1.0 kcal/mol.³⁶ The errors for the differential hydration of conformers should be smaller, and results for several standard cases were shown to be in good accord with experimental data.³⁶ However, E/Z conformers may be particularly challenging owing to the accompanying changes in solute–water hydrogen bonding as compared to simpler cases such as the *gauche/anti* equilibria for 1,2-dihaloethanes.³⁶

The computed $\Delta\Delta G_{\text{hyd}}$ values in Table 4 fall in a relatively narrow range, ± 2 kcal/mol, so the shifts are generally not enough to qualitatively change the direction of the E/Z equilibria. The possible exception is thiourea **6**, for which ΔG appears to be 0 ± 1 kcal/mol in all media.²¹ The expectation from classical electrostatics is that, in the absence

of steric effects, the conformer with the larger dipole moment should have a more negative free energy of hydration. Thus, for most cases in Tables 1 and 2, the E conformer is expected to be better hydrated than the Z conformer. In this regard, the results in Table 4 are mixed. For acetic acid (**1**) the E conformer has a 2.92 D larger dipole moment than the Z form and is better hydrated by 2.39 kcal/mol. This value is significantly smaller in magnitude than estimates of $\Delta\Delta G_{\text{hyd}}$ from a QM/MM study in TIP4P water (-4.8 kcal/mol)⁴¹ and from QM/RISM calculations (-5.2 kcal/mol).⁴² If these values are combined with the G3B3 gas-phase result, the prediction is that (E)- and (Z)-acetic acid are nearly equally populated in water at 298 K or, equivalently, that the Brønsted basicities of the syn and anti lone pairs for acetate ion in water are similar.^{43,44} It should be noted that in dilute aqueous solution at neutral pH, less than 1% of acetic acid is not ionized.

Furthermore, the ester **3** and carbamate **7** pairs also have changes of ca. 3 D in dipole moment, but the E conformer is predicted to be better hydrated by only ca. 0.6 kcal/mol. Previous results for **3** from free-energy perturbation calculations in TIP4P water predicted preferential hydration of the E conformer by 3.0 kcal/mol.¹⁶ Most surprisingly, although the E,Z conformer of the carbonate **9** has a 3.60 D larger dipole moment than the Z,Z conformer, the Z,Z conformer is predicted to be better hydrated by 0.26 kcal/mol. The results for N -methylacetamide **2** also appear to be off the mark. There is general consensus that the E/Z equilibrium for **2** is affected little by hydration,^{11,30} while the GB/SA results favor hydration of the Z conformer by 1.43 kcal/mol. The differential hydration arises predominantly from differences in the GB term. The SA term varies by less than 0.1 kcal/mol for these E/Z equilibria.

The noted discrepancies do not reflect obvious problems with the 1.07*CM1A charges that are used in the GB/SA calculations. The computed dipole moments with these charges mimic the G3B3 results well, e.g., the 1.07*CM1A dipole moments for (E,Z)- and (Z,Z)-**9** are 3.70 and 0.44 D, which are close to the G3B3 values of 3.97 vs 0.37 D. In addition, for (E)- and (Z)- N -methylacetamide, the 1.07*CM1A dipole moments are 4.00 and 3.44 D while the G3B3 results are 4.54 and 4.22 D. Further examination of solvent effects on the E/Z equilibria is warranted using free-energy methods in simulations with explicit solvent. In view of the expected sensitivity of the results to details of solute–solvent hydrogen bonding, it is unclear if continuum models can accurately gauge solvent effects in such cases.

Conclusion

Changes in energy, enthalpy, free energy, and dipole moment were evaluated at the G3B3 level for 18 pairs of conformers exhibiting prototypical E/Z conformational equilibria for rotation about single bonds. The results are important for consideration in molecular design and in evaluation of structures that arise from protein–ligand docking studies as well as from crystallography. For the systems studied, which included representatives of carboxylic acids, carboxylic esters, thioesters, secondary amides, ureas, carbamates, carbonates, enol ethers, enamines, and amidines, the preferred

conformer is normally *Z*. Preference for the *E* conformer mostly arises from steric effects in hydrazones, amidines, and oximes that destabilize the *Z* conformer, especially via *syn*-1,5-CH₃-CH₃ interactions. A particularly interesting case is 1,3-dimethylthiourea, which is found to slightly favor the *E,Z* conformer over the *Z,Z* alternative in the gas phase. Free energies of hydration were also estimated for the conformers from GB/SA calculations. Accurate computation of the effects of hydration on *E/Z* equilibria is expected to be particularly challenging in view of the substantial, accompanying changes in solute-water hydrogen bonding. Though the differential effects from the GB/SA calculations were generally found to be insufficient to overcome the gas-phase preferences, the computed effects in several cases seem too small. Further investigation is warranted with free-energy methods in molecular dynamics or Monte Carlo simulations using explicit hydration to obtain more accurate results and to provide a basis for testing and improvement of continuum solvation methods.

Acknowledgment. Gratitude is expressed to the National Science Foundation and National Institutes of Health (GM32136) for support of this work.

Supporting Information Available: Tables of absolute and relative gas-phase energies, enthalpies, free energies, and dipole moments for **1–18** from the quantum mechanical calculations. This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- Brameld, K. A.; Kuhn, B.; Reuter, D. C.; Stahl, M. Small Molecule Conformational Preferences Derived from Crystal Structure Data. A Medicinal Chemistry Focused Analysis. *J. Chem. Inf. Model.* **2008**, *48*, 1–24.
- Jabs, A.; Weiss, M. S.; Hilgenfeld, R. Non-proline Cis Peptide Bonds in Proteins. *J. Mol. Biol.* **1999**, *286*, 291–304.
- Klebe, G. Virtual ligand screening: strategies, perspectives and limitations. *Drug Discovery Today* **2006**, *11*, 580–594.
- Jorgensen, W. L. Efficient Drug Lead Discovery and Optimization. *Acc. Chem. Res.* **2009**, *42*, 724–733.
- Nichols, S. E.; Domaoal, R. A.; Thakur, V. V.; Bailey, C. M.; Wang, L.; Tirado-Rives, J.; Anderson, K. S.; Jorgensen, W. L. Discovery of Wild-type and Y181C Mutant Non-nucleoside HIV-1 Reverse Transcriptase Inhibitors Using Virtual Screening with Multiple Protein Structures. *J. Chem. Inf. Model.* **2009**, *49*, 1272–1279.
- Ren, J.; Diprose, J.; Warren, J.; Esnouf, R. M.; Bird, L. E.; Ikemizu, S.; Slater, M.; Milton, J.; Balzarini, J.; Stuart, D. I.; Stammers, D. K. Phenylethylthiazolylthiourea (PETT) non-nucleoside inhibitors of HIV-1 and HIV-2 reverse transcriptases: Structural and biochemical analyses. *J. Biol. Chem.* **2000**, *275*, 5633–5639.
- Leach, A. R.; Shoichet, B. K.; Peishoff, C. E. Prediction of protein-ligand interactions. Docking and scoring: successes and gaps. *J. Med. Chem.* **2006**, *49*, 5851–5855.
- Ponder, J. W.; Case, D. A. Force Fields for Protein Simulations. *Adv. Protein Chem.* **2003**, *66*, 27–85.
- Jorgensen, W. L.; Tirado-Rives, J. Potential energy functions for atomic-level simulations of water and organic and biomolecular systems. *Proc. Nat. Acad. Sci. U.S.A.* **2005**, *102*, 6665–6670.
- Wiberg, K. B.; Laidig, K. E. Barriers to rotation adjacent to double bonds. 3. The carbon-oxygen barrier in formic acid, methyl formate, acetic acid, and methyl acetate. The origin of ester and amide resonance. *J. Am. Chem. Soc.* **1987**, *109*, 5935–5943.
- Jorgensen, W. L.; Gao, J. Cis - trans energy difference for the peptide bond in the gas phase and in aqueous solution. *J. Am. Chem. Soc.* **1988**, *110*, 4212–4216.
- Remko, M.; Scheiner, S. The geometry and internal rotational barrier of carbamic acid and several derivatives. *J. Mol. Struct.: THEOCHEM* **1988**, *180*, 175–188.
- Glaser, R.; Streitwieser, A. Configurational and conformational preferences in oximes and oxime carbanions: ab initio study of the syn effect in reactions of oxyimine enolate equivalents. *J. Am. Chem. Soc.* **1989**, *111*, 7340–7348.
- Stang, P. J.; Kitamura, T.; Arif, A. M.; Karni, M.; Apeloig, Y. A single Crystal Structure Determination and Theoretical Calculations on Alkynyl Carboxylate Esters. *J. Am. Chem. Soc.* **1990**, *112*, 374–381.
- Wiberg, K. B.; Wong, M. W. Solvent Effects 4: Effect of solvent on the *E/Z* energy difference for methyl formate and methyl acetate. *J. Am. Chem. Soc.* **1993**, *115*, 1078–1084.
- Evansack, J. D.; Houk, K. N.; Briggs, J. M.; Jorgensen, W. L. Quantification of Solvent Effects on the Acidities of *Z* and *E* Esters from Fluid Simulations. *J. Am. Chem. Soc.* **1994**, *116*, 10630–10638.
- Deerfield, D. W.; Pedersen, L. G. An ab initio quantum mechanical study of thioesters. *J. Mol. Struct.: THEOCHEM* **1995**, *358*, 99–106.
- Sun, H.; Mumby, S. J.; Maple, J. R.; Hagler, A. T. Ab initio calculations on small molecule analogues of polycarbonates. *J. Phys. Chem.* **1995**, *99*, 5873–5882.
- Strassner, T. *Ab Initio* and Molecular Mechanics Calculations of Various Substituted Ureas - Rotational Barriers and a New Parametrization for Urea. *J. Mol. Model.* **1996**, *2*, 217–226.
- Murphy, R. B.; Pollard, W. T.; Friesner, R. A. Pseudospectral localized generalized Moller-Plesset methods with a generalized valence bond reference wave function: Theory and calculation of conformational energies. *J. Chem. Phys.* **1997**, *106*, 5073–5084.
- Chambers, C. C.; Archibong, E. F.; Jabalameli, A.; Sullivan, R. H.; Giesen, D. J.; Cramer, C. J.; Truhlar, D. G. Quantum mechanical and ¹³C dynamic NMR study of 1,3-dimethylthiourea conformational isomerizations. *J. Mol. Struct.: THEOCHEM* **1998**, *425*, 61–68.
- Császár, A. G.; Allen, W. D.; Schaefer, H. F., III. In pursuit of the ab initio limit for conformational energy prototypes. *J. Chem. Phys.* **1998**, *108*, 9751–9764.
- Villani, V.; Alagona, G.; Ghio, C. Ab initio studies on *N*-methylacetamide. Stationary point search and intrinsic reaction coordinate approach. *Mol. Eng.* **1999**, *8*, 135–153.
- Kang, Y. K. Ab initio MO and density functional studies on *trans* and *cis* conformers of *N*-methylacetamide. *J. Mol. Struct.: THEOCHEM* **2001**, *546*, 183–193.
- Senent, M. L. Ab initio determination of the torsional spectra of acetic acid. *Mol. Phys.* **2001**, *99*, 1311–1321.
- Kobychev, V. B.; Larionova, E. Y.; Klyba, N. S. Ab initio study of the conformational and geometric isomerism in

- heteroallyl and heteropropenyl systems. *J. Struct. Chem.* **2003**, *44*, 748–756.
- (27) Zhong, H.; Stewart, E. L.; Kontoyianni, M.; Bowen, J. P. Ab initio and DFT conformational studies of propanal, 2-butanone, and analogous imines and enamines. *J. Chem. Theory Comput.* **2005**, *1*, 230–238.
- (28) Bryantsev, V. S.; Firman, T. K.; Hay, B. P. Conformational Analysis and Rotational Barriers of Alkyl- and Phenyl-Substituted Urea Derivatives. *J. Phys. Chem. A* **2005**, *109*, 832–842.
- (29) Bryantsev, V. S.; Hay, B. P. Conformational preferences and internal rotation in alkyl- and phenyl-substituted thiourea derivatives. *J. Phys. Chem. A* **2006**, *110*, 4678–4688.
- (30) Mantz, Y. A.; Branduardi, D.; Bussi, G.; Parrinello, M. Ensemble of Transition State Structures for the Cis-Trans Isomerization of *N*-Methylacetamide. *J. Phys. Chem. B* **2009**, *113*, 12521–12529.
- (31) Maçôas, E. M. S.; Khriachtchev, L.; Pettersson, M.; Fausto, R.; Räsänen, M. Rotational Isomerism in Acetic Acid: The First Experimental Observation of the High-Energy Conformer. *J. Am. Chem. Soc.* **2003**, *125*, 16188–16189.
- (32) 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.
- (33) Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Rassolov, V.; Pople, J. A. Gaussian-3 (G3) theory for molecules containing first- and second-row atoms. *J. Chem. Phys.* **1998**, *109*, 7764–7774.
- (34) Baboul, A. G.; Curtiss, L. A.; Redfern, P. C.; Raghavachari, K. Gaussian-3 theory using density functional geometries and zero-point energies. *J. Chem. Phys.* **1999**, *110*, 7650–7657.
- (35) Jorgensen, W. L.; Tirado-Rives, J. Molecular modeling of organic and biomolecular systems using BOSS and MCPRO. *J. Comput. Chem.* **2005**, *26*, 1689–1700.
- (36) Jorgensen, W. L.; Ulmschneider, J. P.; Tirado-Rives, J. Free energies of hydration from a generalized Born model and an all-atom force field. *J. Phys. Chem. B* **2004**, *108*, 16264–16270.
- (37) Pawar, D. M.; Khalil, A. A.; Hooks, D. R.; Collins, K.; Elliott, T.; Stafford, J.; Smith, L.; Noe, E. A. E and Z Conformations of Esters, Thiol Esters, and Amides. *J. Am. Chem. Soc.* **1998**, *120*, 2108–2112.
- (38) (a) Huisgen, R.; Ott, H. Medium-sized rings. XV. Configuration of the ester group and the singular properties of lactones. *Tetrahedron* **1959**, *6*, 253–267. (b) Schweizer, W. B.; Dunitz, J. D. Structural Characteristics of the Carboxylic Ester Group. *Helv. Chim. Acta* **1982**, *65*, 1547–1554.
- (39) Nobes, R. H.; Radom, L.; Allinger, N. L. Equilibrium Conformations of Higher-Energy Rotationla Isomers of Vinyl Alcohol and Methyl Vinyl Ether. *J. Mol. Struct.: THEOCHEM* **1981**, *85*, 185–194.
- (40) Yamabe, S.; Tsuchida, N.; Yamazaki, S. Is the Beckmann Rearrangement a Concerted or Stepwise Reaction? A Computational Study. *J. Org. Chem.* **2005**, *70*, 10638–10644.
- (41) Gao, J.; Pavelites, J. J. Aqueous Basicity of the Carboxylate Lone Pairs and C-O Barrier in Acetic Acid: A Combined Quantum and Statistical Mechanical Study. *J. Am. Chem. Soc.* **1992**, *114*, 1912–1914.
- (42) Sato, H.; Hirata, F. The syn-/anti-conformational equilibrium of acetic acid in water studied by the RISM-SCF/MCSCF method. *J. Mol. Struct.: THEOCHEM* **1999**, *461*, 113–120.
- (43) Li, Y.; Houk, K. N. Theoretical Assessments of the Basicity and Nucleophilicity of Carboxylate Syn and Anti Lone Pairs. *J. Am. Chem. Soc.* **1989**, *111*, 4505–4507.
- (44) Rebek, J., Jr. Molecular Recognition with Model Systems. *Angew. Chem., Int. Ed.* **1990**, *29*, 245–255.

CT1004017

Understanding the Mechanism for Ribonucleotide Reductase Inactivation by 2'-Deoxy-2'-methylenecytidine-5'-diphosphate

M. A. S. Perez, P. A. Fernandes, and M. J. Ramos*

REQUIMTE, Faculdade de Ciências, Universidade do Porto, Rua do Campo Alegre, 687, 4169-007 Porto, Portugal

Received April 24, 2010

Abstract: Ribonucleotide reductase (RNR) is the key enzyme in the biosynthesis of deoxyribonucleotides. The enzyme has thus an attractive target for chemotherapies that fight proliferation-based diseases. 2'-Deoxy-2'-methylenecytidine-5'-diphosphate (CH_2dCDP) is a potent mechanism-based inhibitor of the enzyme RNR, which decomposes to an active alkylating furanone specie. The details of the inhibition mechanism are unknown, and experimental studies have indicated that some properties of the inactivation are dissimilar to those observed with a number of 2'-substituted 2'-deoxynucleotides mechanism-based inhibitors. To disclose the mechanism involved in RNR inactivation by CH_2dCDP we explored the potential-energy surface in two different models of the system with different objectives in mind. In order to conveniently explore the reactional space, i.e. to study the possible reactions between the CH_2dCDP and the RNR, we used a small model representing the active site of RNR with CH_2dCDP using DFT. To provide further insights and efficiently account for the long-range RNR– CH_2dCDP interactions and the stereochemical strain imposed by the protein scaffold we performed theoretical calculations on the more promising reactions using hybrid QM/MM calculations on a larger model system. We used quantum mechanics for the active-site region (CH_2dCDP and active-site residues) and molecular mechanics for the surroundings (6373 atoms of the R1 monomer). The results obtained led us to understand the correct mechanism for RNR inactivation by CH_2dCDP , and the furanone species formed presumably explains the dissimilarities observed with a number of 2'-substituted 2'-deoxynucleotides.

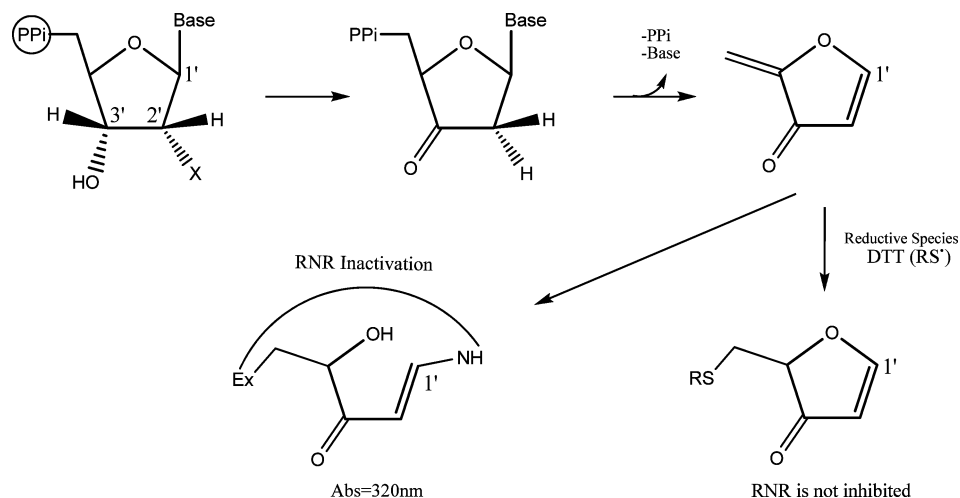
Introduction

The reduction of ribonucleotides into deoxyribonucleotides is strictly conserved in all living organisms and catalyzed by ribonucleotide reductase (RNR). This key role makes the RNR involvement a rate-limiting step in DNA replication and repair, turning it into an attractive target for antitumor, antiviral, and antibacterial therapies.^{1–8}

Different organisms have related RNRs, and their catalytic mechanism is always based on radical-mediated reactions.⁹ The *Escherichia coli* (*E. coli*) RNR is similar to the mammalian RNR and has served as its prototype in experimental and theoretical studies. *E. coli* RNR is consti-

tuted by two different homodimeric subunits. The larger one, known as R1 (761 residues), lodges the active site (for reduction of purines and pyrimidines) and three independent allosteric sites (specificity site, adenine specific site and hexamerization site).^{10,11} The smaller subunit, known as R2 (375 residues), contains a tyrosyl radical that is stabilized by an oxo-bridged binuclear Fe(III) complex.^{12,13} The active-site residues that are catalytically important are Cys225, Cys439, Cys462, Glu441, and Asn437. For catalysis to take place, the tyrosyl radical in R2 must generate a thyl radical in Cys439 of the R1 active site. The catalysis can only occur when the tyrosyl radical migrates from subunit R2 to subunit R1 active site. The electron migration is thought to occur by proton-coupled electron transfers, through a ca. 35 Å long

* Corresponding author e-mail: mjramos@fc.up.pt.

Scheme 1. RNR Inactivation by 2'-Halo-2'-deoxynucleoside-5'-diphosphate²⁴

chain of hydrogen-bonded residues.^{14–20} The substrates of RNR are ribonucleosides-5'-diphosphates.

Different strategies for inactivation of RNR have been reported, namely, several modifications of the nucleotides' ribose moiety at the 2' and 3' positions have produced potent mechanism-based inhibitors. 2'-Deoxy-2'-methylencytidine-5'-diphosphates (*CH*₂dNDPs) are included among them.²¹ 2'-Deoxy-2'-methylidene-nucleosides (*CH*₂dN) were synthesized by Takenuki et al.²² in 1988 and have been demonstrated to possess antitumor activity.²³ It has been found that 2'-deoxy-2'-methylideneuridine-5'-diphosphates (*CH*₂dUDP) and 2'-deoxy-2'-methylidencytidine-5'-diphosphates (*CH*₂dCDP) function as irreversible inactivators of RNR, but *CH*₂dCDP is a much more potent inhibitor. Experimental studies have shown that the inactivation is initiated by carbon C3'–H bond cleavage resulting in a 3'-keto-2'-methyldeoxyribose nucleotide as the end product.²¹ The collapse of this into a furanone derivative occurs only in solution, later becoming attached to protein R1 and inhibiting the enzyme. The inhibition process is detected experimentally by a rapid increase in absorbance at 326 nm, and the rate of appearance of this band is similar to the rate of inactivation. However, this absorption band disappears shortly afterward, and a new broad absorption band is detected at 366 nm with a substantially decreased extinction coefficient. The release of cytosine and protection against enzyme inactivation by reductive species are very similar to results observed with a number of substrate analogues; however, the rate of increase in absorbance at 326 nm and the subsequent decrease are considerably different from any of the previously studied compounds.^{2,21} The rate of inactivation by 2'-halo-2'-deoxynucleoside-5'-diphosphates is fast under identical conditions, and the change in absorbance on the protein at 320 nm is very slow. These results have been interpreted as indicating rapid inactivation by a nucleophilic attack of a group on the enzyme at the more reactive exocyclic methylene of 2-methylene-3(2*H*)-furanone.²⁴ The absorbance increase at 320 nm results from a subsequent attack of a lysine residue at the C1' position, ultimately leading to the putative α,β -unsaturated enamine (Scheme 1). The kinetics of the formation and the decay observed in the inhibition by *CH*₂dCDP has not yet been unravelled.

Taking into account all experimental data and earlier mechanistic studies with other 2'-substituted-nucleotide analogues, theoretical calculations on two different models of the system were performed in order to thoroughly explore the RNR:*CH*₂dCDP potential-energy surface (PES) and consequently understand the mechanism for RNR inactivation by *CH*₂dCDP. The rate of inactivation presumably can be explained by the furanone species formed.

Computational Details

Small Model. We explored the possible reactions between *CH*₂dCDP and RNR with a small model representing the active-site residues of RNR (Cys439, Cys462, Cys225, and Glu441) with *CH*₂dCDP. In this model, cysteines and glutamate residues were built as methylthiols and formate molecules, respectively. The *CH*₂dCDP molecule was modeled without the base and the diphosphate. All calculations were performed using density functional theory (DFT), with the Gaussian03 suite of programs,²⁵ at the unrestricted Becke3LYP (B3LYP) level of theory^{26–28} with the 6-31G(d) basis set. The adequacy of these models, level of theory, and basis set were demonstrated in earlier results.^{29–32} Frequency calculations confirmed the nature of each stationary point, i.e., an energy minimum with all frequencies real or, in the case of a transition state, one imaginary frequency only. The transition states were verified to connect the reactants and products of interest through internal reaction-coordinate calculations (IRC). Zero-point, thermal, and entropic effects ($T = 298.15$ K, $P = 1$ bar) were added to the calculated energies. A scale factor of 0.9804 was used to eliminate known systematic errors in zero-point energies and thermal-energy corrections. The atomic spin-density distributions were calculated by employing a Mulliken population analysis.²⁵ All optimized energies were corrected by single-point energy calculations with the 6-311++G(2d,2p) basis set and a polarized-continuum model. The polarized-continuum model, C-PCM, was employed as implemented in Gaussian03.²⁵ In this approach a continuum is modeled as a conductor instead of a dielectric. This simplifies the electrostatic computations, and corrections are made a posteriori for dielectric behavior. A dielectric constant of

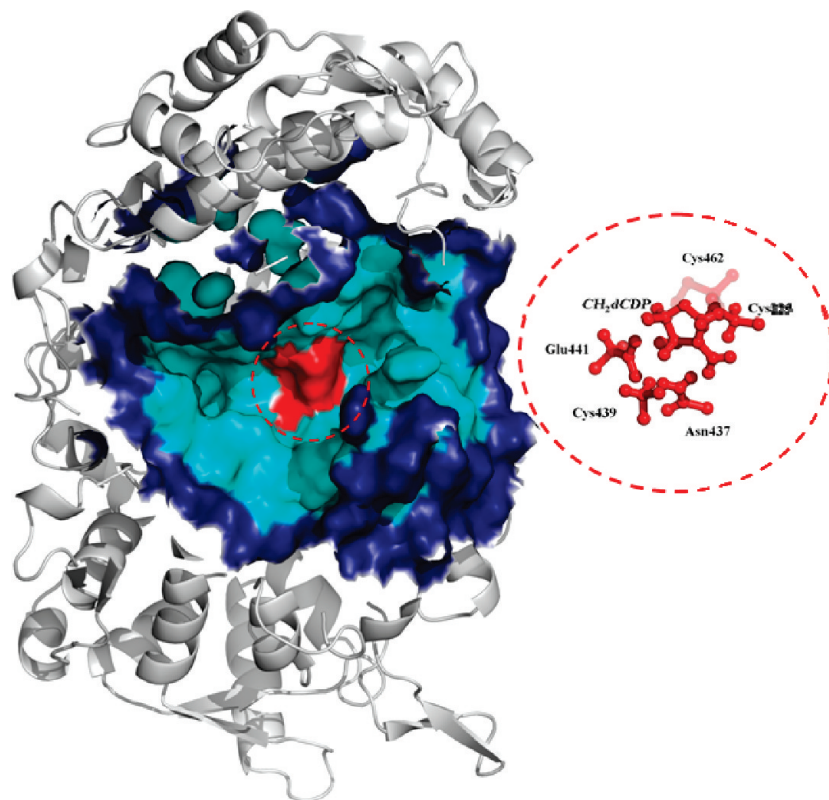


Figure 1. Representation of the enzymatic model studied, which includes a 20 Å radius of the amino acids around CH_2dCDP . The higher level region used in the QM/MM calculations is colored red, and the region treated with the lower level is colored cyan (the free region) and marine (the frozen region, the outer 5 Å shell of the entire system). The excluded part of the R1 monomer is shown in gray.

four³³ has shown previously to provide theoretical results that are in agreement with experimental results and accounts for the combined effect of the protein and buried water molecules.

Large Model. The R1 protein complexed with the inhibitor CH_2dCDP was modeled based on the crystal structure 4R1R, ribonucleotide reductase R1 protein with substrate, GDP, and effector DTTP from *E. coli*, determined by Eriksson et al.³⁴ Just one of the R1 monomers was used in the calculations. Taking into account that RNR is highly specific for ribonucleotides and they fit very tightly into the three binding pockets, (the ribose, phosphate, and base binding pockets) we created the inhibitor CH_2dCDP using the substrate as reference. The hydrogen atom and hydroxyl group that were bound to the carbon C2' of the substrate were replaced by a methylene. For the modifications we used the GaussView²⁵ software. To calculate the molecular electronic structures and properties for CH_2dCDP and DTTP we used Gaussian03,²⁵ performing restricted Hartree–Fock calculations (RHF), with the 6-31G(d) basis set, to be consistent with the parametrization adopted in Amber 8,³⁵ which was later used in the molecular mechanics calculations. We used the Antechamber tools within Amber 8, in order to create an input file for CH_2dCDP and DTTP that could be read by Leap, another of Amber's modules. Antechamber is designed to be used with GAFF, the general amber force field. RESP³⁶ was the method used to calculate atomic charges.

The R1 subunit complexed with the CH_2dCDP was minimized in order to release the bad contacts. The minimizations were performed with the parametrization adopted in Amber 8,³⁵ using the Amber force field ff99 for proteins and the Gaff force field for CH_2dCDP and DTTP.^{37–39} We solvated our complex with an octahedral box of water, using an 8 Å buffer of TIP3P water model. Our minimization procedure for solvated R1 subunit complexed with CH_2dCDP consists of a two-stage approach. In the first stage we kept the complex fixed and minimized the positions of the water and ions. Then in the second stage we minimized the entire system. About 5000 steps were used for each stage, with the first 2500 steps performed using the steepest descent algorithm and the remaining steps carried out using the conjugate gradient algorithm.

To perform the study, we used the final structure of the minimization, from which we built a model including a 20 Å radius of the enzyme around the CH_2dCDP molecule (Figure 1).

The model system was composed by a total of 6373 atoms. The QM/MM calculations performed to determine the potential-energy surface (PES) of the enzymatic model system were made with the Gaussian 03 software.²⁵ To explore the PES of the catalytic reaction, the system was divided into two layers within the ONIOM formalism^{40,41} as implemented in Gaussian 03. The higher level layer included the inhibitor without the base and the diphosphate and the side chains of the residues Cys439, Cys225, Cys462, Glu441, and Asn437 in a total of 49 atoms (Figure 1). The

higher layer was treated with density functional theory (DFT) at the unrestricted B3LYP/6-31G(d) level.^{26–28} The rest of the system was treated at the molecular mechanics level with AMBER. We have further frozen the positions of the atoms in the outer 5 Å shell of the complex (Figure 1).

For each reaction step, we performed a linear transit scan along the reaction coordinate with a step value of 0.05 Å in order to locate the geometry of the transition state. We considered the higher energy geometry as a very good approximation to the geometry of the transition state. We used the electronic embedding²⁵ scheme during the scan. Electronic embedding incorporates the partial charges of the MM region into the QM calculations and allows for polarization of the QM region by the MM charges. This technique provides a better description of the electrostatic interaction between the QM and the MM regions. The atomic spin density distributions were calculated with a Mulliken population analysis²⁵ in the higher level calculation. We calculated the energy of the optimized geometries of reactants, products, and transition states for each mechanistic step. The inclusion of diffuse functions in the basis set for geometry optimizations was investigated before and concluded that the corrections to the geometry were very small and corrections in energy differences (activation and reaction energies) were negligible.⁴² We therefore considered it unnecessary from a computational point of view to include diffuse functions in geometry optimizations, considering the inherent increase in computing time. Single-point energy calculations were then performed on the optimized geometries of the reactants, transition state, and products increasing the basis set of the higher level region to 6-311++G(2d,2p). Single-point energy calculations were also performed on all optimized geometries with the BB1K functional^{26,27,43} and two different basis sets (6-31G(d) and 6-311++G(2d,2p)). In order to understand more precisely the relative importance of long-range electrostatic/polarization effects, additional single-point energy calculations were performed on all optimized QM/MM geometries with the mechanical embedding scheme (ONIOM-ME) treating the higher level with the BB1K functional and the 6-311++G(2d,2p) basis set. ONIOM-ME neglects any polarization effects of the MM charges in the QM region.

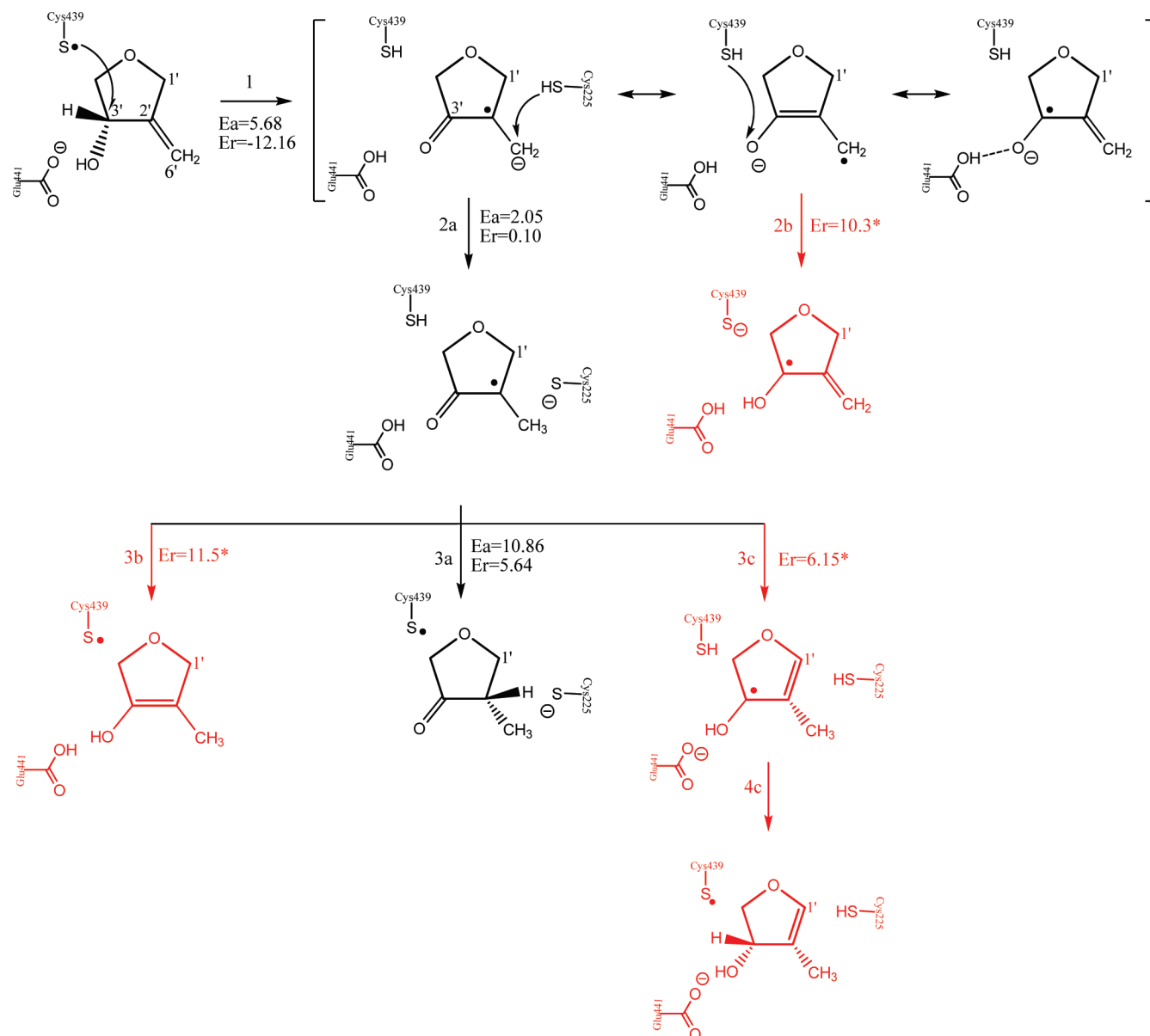
Results

Considering earlier mechanistic results with other 2'-substituted substrate analogues we tried to explore all relevant mechanistic pathways with the small model. Scheme 2 shows an overview of the reactions studied with the small model. The activation and reaction energies presented in Scheme 2 have been obtained with B3LYP using the 6-311++G(2d,2p) basis set and a polarized continuum model. Zero-point energies, thermal, and entropic effects obtained from B3LYP/6-31G(d) were added.

The first step involves abstraction of a hydrogen from carbon C3' of *CH*₂dCDP by radical Cys439 and is known to occur in the natural catalysis and with most of the substrate analogues, leading to a stable product with the spin density and the negative charge delocalized in the inhibitor and with residue Cys225 nearer to the inhibitor, more precisely closer

to the C6' carbon. The first step is very exothermic, and the reaction would be practically irreversible ($E_a = 5.68$ kcal/mol and $E_r = -12.16$ kcal/mol). The energetics of this abstraction are dissimilar to the ones obtained in an earlier study with the substrate ($E_a = 11.09$ kcal/mol, $E_r = 8.01$ kcal/mol) but more similar to that observed in earlier studies with substrate analogues, namely, a study with the substrate analogue *CHF*dCDP, which has a fluormethylene in the 2' position ($E_a = 8.01$ kcal/mol and $E_r = -7.7$ kcal/mol).⁴⁴ Taking into account the characteristics of the product of step 1 and the position of the conserved residues around, steps 2a and 2b are more promising. The product of the barrierless step 2b lies about 10 kcal/mol above the reactant and ends up falling back into it, indicating that it does not correspond to a stationary point in the potential-energy surface. There is no doubt that, out of the two, step 2a is preferential because it has a very small energy barrier ($E_a = 2.05$ kcal/mol) and is athermic ($E_r = 0.10$ kcal/mol). At this point, the inhibitor has the radical essentially located on the C3' atom and the negative charge essentially located in residue Cys225 and there are two steps that lead to regeneration of the radical sulfur of the Cys439, steps 3a and 3b, and one step, step 3c, which seems relevant due to the proximity between carbon C1' of the inhibitor and residue Cys225. The products of steps 3b and 3c do not correspond to stationary points on the potential-energy surface, and they fall back into the reactant. It seems logical that step 3c is unfavorable comparatively with step 3a; however, we decided to confirm with the larger model because the energy required to form the unstable compound is not very high and step 4c probably could stabilize it. The reactions (steps 1, 2a, 3a, and 3c) were tested with the larger model. The larger model allowed us to discard step 3c beyond any doubt; the energy barrier obtained for this step (more than 35 kcal/mol) is too high to match experimental kinetics. The large difference between the small and large model is unique in this step. This is due to the large stereochemical constraints imposed by the protein for Cys225 to reach C1. In the small model those constraints are absent. No other step is expected to exhibit a similar difference between models. We note that each transition state of each step is characterized by a unique imaginary frequency, specifically, 1291i cm⁻¹ for step 1, 725.3i cm⁻¹ for step 2a, and 1072i cm⁻¹ for step 3a. Finally, the more viable mechanism is shown in detail (Scheme 3).

For the sake of comparing mechanisms B3LYP is a very efficient tool. It is much faster and has better convergence properties than the modern hybrid-meta functionals, two requisites very helpful to explore mechanistic pathways. However, to obtain final values for the PES we preferred to use the BB1K energies, and these are the ones which will be discussed above. BB1K has been shown to significantly outdo B3LYP in several properties, namely, provide more realistic barriers. We also calculated the B3LYP energies for the large model in the preferred pathway and confirmed that in all reactions the activation energies provided by B3LYP (Supporting Information) were up to 1.0–4.5 kcal/mol smaller than the BB1K values. Each mechanistic step will be discussed, presenting the results obtained with the larger model that provides further insights and efficiently

Scheme 2. Overview of the Relevant Reactions Explored with the Small Model in Order To Unravel the Mechanism for RNR Inhibition by CH_2dCDP^a 

^a All energy values are in kcal/mol. The asterisk (*) identifies the cases for which the energy of the products do not correspond to a stationary point in the potential-energy surface but represents the higher energy value of the scan.

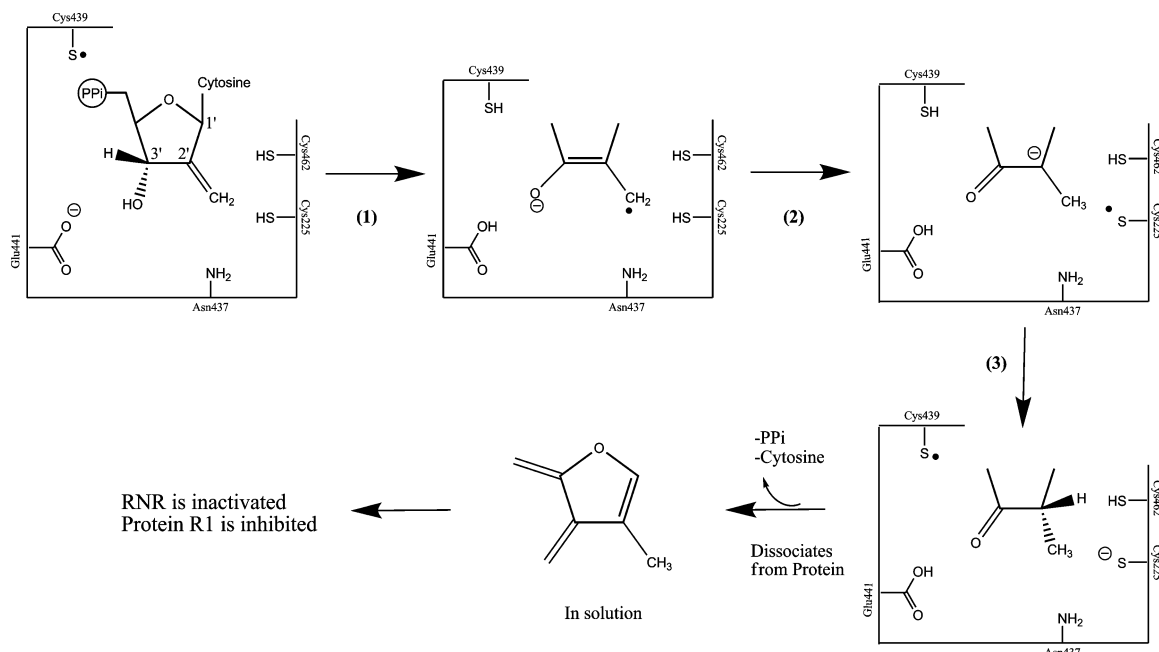
account for the long-range RNR– CH_2dCDP interactions and stereochemical strain imposed by the protein scaffold. In order to calculate the activation and reaction energies, the energy of each state has been calculated as the energy obtained with ONIOM (using BB1K/6-311++G(2d,2p) for the higher layer and AMBER for the rest of the system), corrected by zero-point energy, thermal, and entropic effects (eq 1).

$$E = E_{\text{ONIOM(BB1K/6-311++G(2d,2p)//AMBER)}} + (\text{ZPE} + \text{thermal} + \text{entropic})_{\text{B3LYP/6-31G(d)}} \quad (1)$$

First Mechanistic Step. The first step involves abstraction of a hydrogen from carbon C3' of CH_2dCDP by radical Cys439 previously formed, as observed with the natural

substrate and with some substrate analogues previously studied. Similarly, one proton of the hydroxyl group connected to the C3' of the inhibitor ($HO-C3'$) migrates to residue Glu441 spontaneously. This rearrangement of the system is energetically and thermodynamically favorable, the calculated free energy barrier necessary to achieve the transition state is 6.39 kcal/mol, and the reaction free energy is -22.08 kcal/mol.

In the reactants, the thiol atom of Cys439, where the spin density is located (0.91 au), is found at 2.60 Å from the hydrogen that is bound to the C3' carbon ($H-C3'$) and the proton atom of the hydroxyl group connected to the C3' ($HO-C3'$) is at 1.60 Å from the oxygen atom of the Glu441. Figure 2 represents the geometry of the transition state. At this point the hydrogen is located 1.65 Å away from the

Scheme 3. Proposed Mechanism for RNR Inhibition by CH_2dCDP 

sulfur (Cys439) and 1.32 Å from the carbon C3' of the inhibitor. The proton of the HO-C3' group is in an almost perfectly collinear fashion with the oxygen atom of the Glu441 with an O-H-O angle of 177.5° and is still closer to the O-C3' bond (1.10 Å from the oxygen atom that is bound to the C3' carbon and 1.40 Å from the oxygen of the Glu441). The C3'-O and C3'-C2' bonds are 0.05 and 0.06 Å shorter than in the reactants, and the C3'-C6' bond is 0.10 Å elongated. In the optimized geometry of the products the C3'-O and C3'-C2' bonds are further shortened to 1.30 and 1.41 Å, respectively, and the C3'-C6' bond is more elongated to 1.38 Å. The spin density is essentially located in the C6' atom (0.61 au) but delocalized to the carbon C3' (0.36 au), and the negative charge of the system becomes localized in the inhibitor because Glu441 is now protonated.

Second Mechanistic Step. At the end of step 1, the C2'-C6' bond has become elongated and the spin density is essentially located in the C6' atom, revealing that C6' can easily receive one hydrogen. The hydrogen atom of Cys225 is distanced by 2.43 Å from the C6' atom of the inhibitor. The second step consists of the transfer of a hydrogen from Cys225 to the carbon C6' of the inhibitor. The transition state for reaction 2 is represented in Figure 3.

At this point the hydrogen is shared between the sulfur (1.62 Å) of Cys225 and the carbon C6' of the inhibitor (1.45 Å). The distance C2'-C6' is further elongated to 1.42 Å, and the spin density is essentially in carbon C6' (0.36 au) and in the sulfur atom of residue Cys225 (0.28 au). In the optimized structure of the products, the sulfur atom of

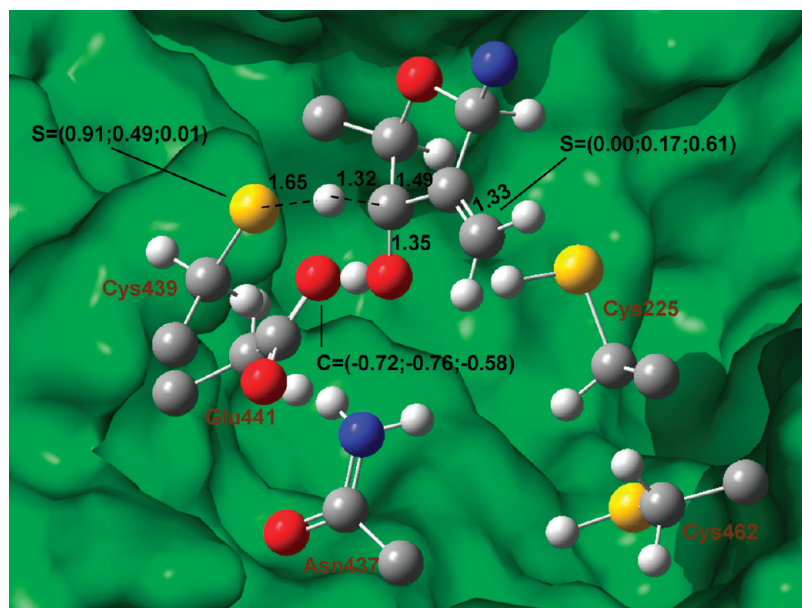


Figure 2. Geometry of transition state 1 labeled with relevant bond lengths (in Angstroms), spin density distribution (S) in au, and the charges (C) of the three geometries (reactants, transition state, and products).

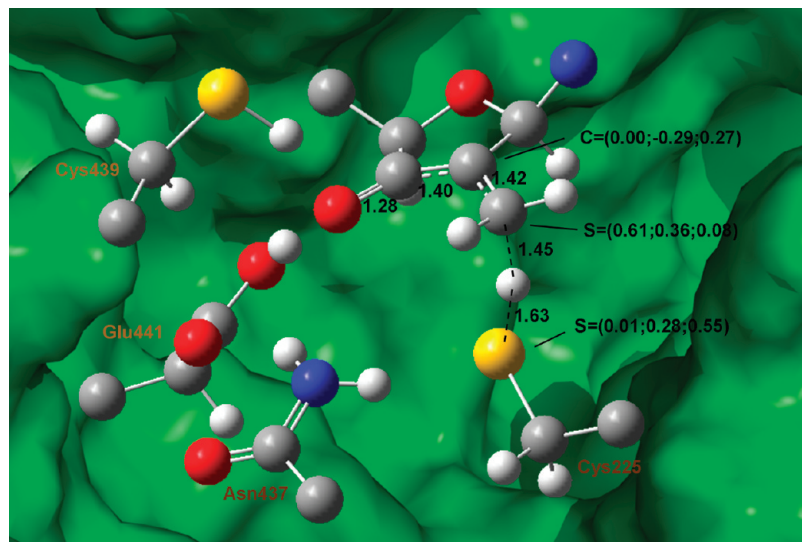


Figure 3. Geometry of transition-state 2 labeled with relevant bond lengths (in Angstroms), spin density distribution (S) in au, and charges (C) of the three geometries (reactants, transition state, and products).

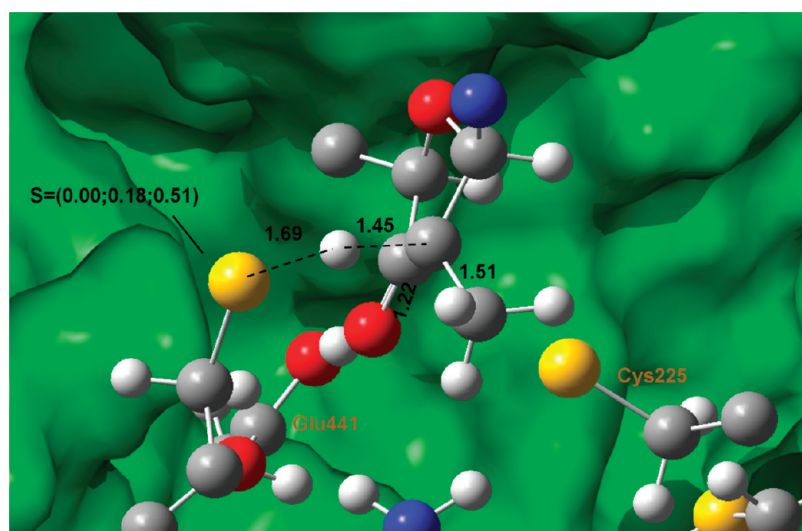


Figure 4. Geometry of transition-state 3 labeled with relevant bond lengths (in Angstroms), spin density distribution (S) in au, and the charges (C) of the three geometries (reactants, transition state, and products).

Cys225 is 2.43 Å away from the hydrogen, the C6'–C2' bond is 1.47 Å and the C2'–C3' and C3'–O bonds are shorter (1.38 and 1.29 Å, respectively). The spin density is mainly located at the sulfur atom of the Cys225 (0.55 au), although it is still delocalized to carbon C2'. As a consequence of resonance, the negative charge is delocalized between the C2' of the inhibitor and the oxygen atom that is bound to the C3'. The calculated free energy barrier necessary to achieve the transition state is 10.05 kcal/mol, and the reaction free energy is 8.27 kcal/mol.

Third Mechanistic Step. At this point of the reaction pathway, there is clearly one step that would lead to regeneration of the R1 radical. It consists in the donation of the thiol hydrogen of Cys439 to carbon C2'. In the optimized structure of the reactants, the hydrogen of Cys439 is distanced by 3.37 Å to the C2' carbon of the inhibitor. The activation free energy for this reaction is 16.70 kcal/mol, and the reaction free energy is 3.87 kcal/mol. Figure 4 depicts the geometry of the transition state, where the hydrogen is

1.64 Å away from the sulfur atom of Cys439 and 1.45 Å from the carbon C2' at an angle of 174.7°.

In the optimized structure of the products the sulfur atom of Cys439 is 2.43 Å away from the hydrogen, the C6'–C2' bond length is 1.53 Å, C2'–C3' is 1.51 Å, and C3'–O is 1.22 Å. The spin density is essentially in the sulfur atom of the Cys439 (0.51 au), and the negative charge is essentially in the sulfur atom of the Cys225. This way, the initial thiyl radical is regenerated and a 2'-methyl-3'-ketodeoxyribonucleotide is formed. This step has the higher activation energy (also in small model). It must be noted that a barrier of 15.64 kcal/mol is in agreement with the experimental kinetics for this reaction (the barrier for the limiting step is less than 20 kcal/mol).²⁴

RNR Inhibition. According to experimental results, the 2'-methyl-3'-ketodeoxyribonucleotide does not remain attached to the active site and has a propensity to dissociate to the solvent. Once in solution, the 2'-methyl-3'-ketodeoxyribonucleotide loses the base and the phosphate groups and

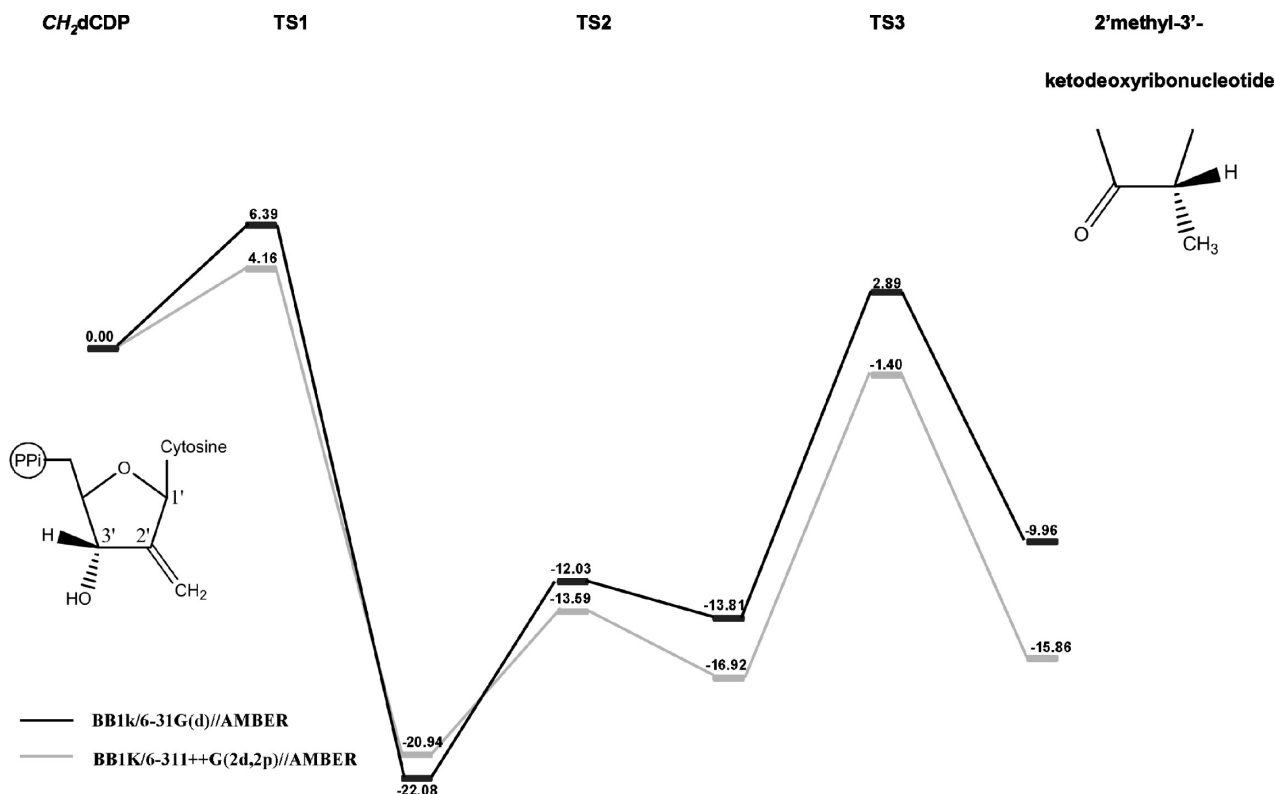


Figure 5. Energetic profile of the mechanism proposed for the RNR interaction with CH_2dCDP .

generates a methylfuranone derivative (Scheme 3). Subsequently, this compound inhibits subunit R1, forming a chromophore with a characteristic absorption band at 326 nm. This seems to be consistent with production of a methyl-substituted α,β -unsaturated enamine by analogy with 2'-halo-2'-deoxynucleoside-5'-diphosphates chemistry (Scheme 1). Replacement of H by CH_3 in the furanone would be expected to red shift the wavelength λ by 5–10 nm.²¹ What is atypical is that the rate of appearance of λ at 326 nm is fast and similar to the rate of inactivation. The rate of inactivation by 2'-halo-2'-deoxynucleoside-5'-diphosphates is fast, and the change in absorbance on the protein at 320 nm is very slow. Probably, the presence of a methyl group bound to the C2' atom would conduct a faster attack of a lysine residue at the C1' position. We searched hard, computationally, but have not found another favorable pathway that would lead to a second chemical species besides the above-mentioned furanone and consequently justify the second absorption value at 366 nm. However, since the absorbance at 326 nm disappears completely and a new absorption band is formed, subsequent chemical rearrangements within the R1 active site might explain this observation. In fact, formation of a single methylfuranone and its subsequent rearrangement would be in agreement with the experimental findings. These rearrangements might be due to a different binding pose/binding location due to the extra methyl group.

Overall Reaction Energy. In Figure 5 we present the overall energetic profile of the mechanism proposed for the RNR interaction with CH_2dCDP now calculated with the more accurate and time-extensive BB1K density functional.

According to Figure 5, the mechanism proposed for RNR inhibition has an overall reaction energy of -15.86 and

-9.96 kcal/mol using the functional BB1K with 6-31G(d) and 6-311++G(2d,2p), respectively, and is therefore thermodynamically favored. We note the functional BB1K overestimates slightly (~ 1 kcal/mol) the energy for the generality of the theoretical studies made with different enzymes.⁴⁵ When comparing the basis sets the activation and reaction energies have augmented in all steps except the first. This would be expected since all steps involve transfer of hydrogens or protons. The first mechanistic step is the most favorable thermodynamically, and the third step is the limiting one, with the higher activation barrier. Note that the overall PES is slightly endothermic after step 1, which would suggest that the reaction could be trapped in the first intermediate. The thermodynamic driving force for the inhibition comes from the very exothermic dissociation of the 2'-methyl-3'-ketodeoxyribonucleotide to solution,⁴⁶ which makes the overall process very favorable. There is no doubt that computationally this is the most viable mechanistic pathway.

Electrostatic/Polarization and Strain at the Active Site. In the proposed mechanism for RNR interaction with CH_2dCDP the ONIOM-EE calculations were used to take into account the long-range interactions, the stereochemical strain imposed by the protein scaffold, and the polarization/electrostatic effect on the QM layer due to the MM layer. In this subsection of the results we comment more precisely on the relative importance of the long-range electrostatic/polarization and stereochemical strain.

To calculate the contribution of the stereochemical strain for the activation and reaction energies we compared the energies of the small free system in the gas phase (ΔE^{free}) with the energies of that same system also in the gas phase

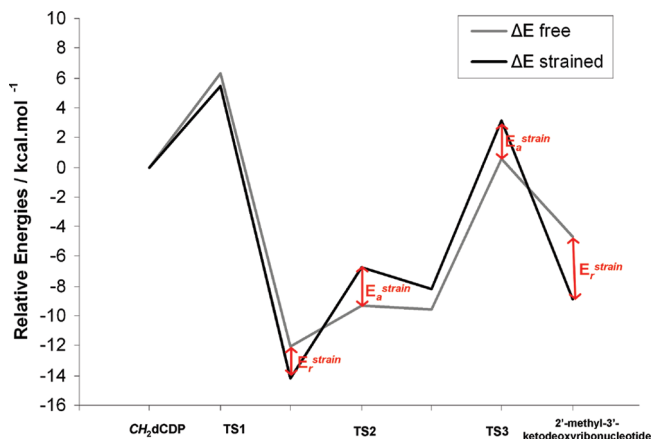


Figure 6. Stereochemical strain contribution for the energy profile.

but with the precise geometry obtained in the full QM/MM model ($\Delta E^{\text{strained}}$). The energies were recalculated to eliminate the effects of the dielectric continuum for the first case and the polarization by the MM region in the second case. The strain contribution for the activation energy ($E_a^{\text{strain}} = \Delta E_a^{\text{strained}} - \Delta E_a^{\text{free}}$) and for the reaction energy ($E_r^{\text{strain}} = \Delta E_r^{\text{strained}} - \Delta E_r^{\text{free}}$) are shown in Figure 6. We note that the driving force that leads to the strained geometry in the QM/MM model is not only the “mechanical strain” imposed by the almost rigid bonds and angles that connect the residues to the backbone but also the electrostatic field created by the remaining protein scaffold. It is obvious that most of the contribution comes from the mechanical strain.

In Figure 6 we can see that the strain energy favors transition-state 1 (TS1) and product 1 (P1), with $E_a^{\text{strain}} = -0.91$ kcal/mol and $E_r^{\text{strain}} = -2.17$ kcal/mol. Strain facilitates the first step. The constraints of the position/orientation of the CH_2dCDP and the reactive-site residues due to their connections to protein scaffold disfavors generation of TS2 and TS3 (E_a^{strain} is ~ 2.5 kcal/mol in both) and P2 ($E_r^{\text{strain}} = 1.37$ kcal/mol). The strain contribution for the overall reaction energy is -4.24 kcal/mol, i.e., the strain due to the protein scaffold favors formation of the 2'-methyl-3'-ketodeoxyribonucleotide. Looking to the overall energetic profile we can see that the constraints imposed by the protein scaffold play a minimal catalyst role in the catalytic mechanism. This is one more argument that supports the

previous use of cluster models in studies of the catalytic mechanism of RNR.^{31,42,47}

An interesting technical issue is the extent to which the electronic polarization of the substrate and active site contributes to the potential-energy profile. The polarization always increases the electronic energy when compared to the gas-phase unpolarized system. However, this energetic cost must be compensated by the increased electrostatic interactions between the polarizing (MM layer) and polarized (QM layer) regions. The contribution of the polarization energy to the activation energy (E_a^{pol}) and reaction energy (E_r^{pol}) is given by the differences in the high-level layer PES calculated with ONIOM-EE and ONIOM-ME ($E_a^{\text{pol}} = \Delta E_a^{\text{high, ONIOM-EE}} - \Delta E_a^{\text{high, ONIOM-ME}}$ and $E_r^{\text{pol}} = \Delta E_r^{\text{high, ONIOM-EE}} - \Delta E_r^{\text{high, ONIOM-ME}}$), always using the same ONIOM-EE geometries. See Figure 7.

The contribution of the polarization energy to TS1 is negligible ($E_a^{\text{pol}} = 0.05$ kcal/mol). Polarization becomes in fact significant after the first step. The long-range polarization effects disfavor TS2 ($E_r^{\text{pol}} = 9.01$ kcal/mol) and favors to a great extent P2, TS3, and the final product. The polarization contribution for the overall reaction energy is -17.25 kcal/mol, extensively favoring formation of the final product 2'-methyl-3'-ketodeoxyribonucleotide. The polarization effect has an important contribution to the stabilization, favoring quantitatively the products of the proposed mechanism. However, analyzing the high-layer polarization effects alone may be misleading, as polarization gives rise to enhanced electrostatic interactions with the MM region, and these two effects should be analyzed together. The enzyme rotameric response to the field created by the substrate is only poorly measured, even in long MD simulations. Atomic polarization at the MM layer is also not explicitly accounted for with the usual biomolecular force fields. The most straightforward way to account for polarization and electrostatic effects together is to compare the results obtained with the electronic embedding and mechanical embedding schemes. We have done so for the first step of the reaction, as an example. However, there is a strong bias in the process, which is the fact that the mechanical embedding results depend on the atomic point charges used to parametrize the MM region. Point charges are usually first calculated by fitting to the QM electrostatic potential (conventional point charges) and

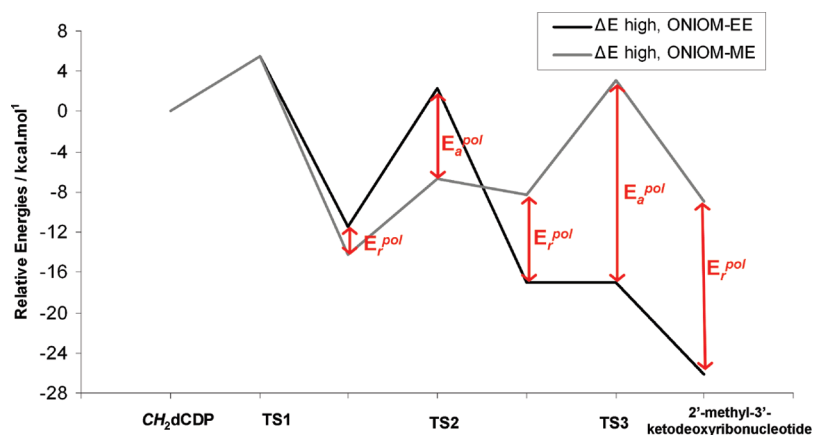


Figure 7. Polarization contribution for the energy profile.

then empirically adjusting to reproduce known data (effective point charges) to overcome limitations inherent to the fitting process and lack of MM atomic polarization. However, the only reference we have is the electronic embedding result, and to fit the charges to reproduce the electronic embedding result would make the assessment of polarization/electrostatic effects biased and meaningless. Therefore we used just conventional point charges (both Mulliken charges and HF/6-31G(d) ESP fit charges) to recalculate the energy of the first step with mechanical embedding. The results for the first step were $E_a = 12.9$ kcal/mol and $E_r = -22.7$ kcal/mol with Mulliken charges and $E_a = 24.0$ kcal/mol and $E_r = -4.56$ kcal/mol with HF/6-31G(d) ESP fit charges, which clearly shows that the results of mechanical embedding are very much dependent on the scheme used to generate point charges and not very reliable (if electrostatic embedding is taken as a reference) unless one has a way to empirically adjust the point charges of the MM layer. Note that inclusion of ESP charges derived from the electrostatic embedding electronic density (i.e., “polarized point charges”) followed by MM geometry reoptimization led to very similar results to the ones obtained with the ESP “unpolarized” charges, which highlights the importance of the parametrization of the QM regions and the dependence of the results on this process. We consider it to be preferable to use electronic embedding as it is more objective and less dependent on the exact choice of the method/protocol to parametrize the high layer.

Conclusions

This study has allowed for comprehension of the mechanism in which 2'-deoxy-2'-methylencytidine-5'-diphosphate inhibits RNR, with atomistic detail. Even though this work was derived from theoretical calculations, all earlier experimental results were taken into account. We propose the more viable mechanistic pathway for RNR inactivation by CH₂dCDP in three steps. The first mechanistic step involves abstraction of a hydrogen from carbon C3' of CH₂dCDP by radical Cys439. In fact and spontaneously, one proton of the hydroxyl group connected to the C3' carbon atom of the inhibitor migrates to Glu441. The second step consists of transfer of a hydrogen from Cys225 to the carbon C6' of the inhibitor. The third step consists of donation of the thiol hydrogen of Cys439 to carbon C2'. The 2'-methyl-3'-ketodeoxyribonucleotide formed does not remain attached to the active site, dissociating instead to the solvent. Once in the solvent, it loses the base and the phosphate groups and generates a methylfuranone derivative. The methylfuranone derivative inhibits subunit R1. Probably, the presence of a methyl group bound to the C2' carbon atom will conduct to a faster attack of a lysine residue at the C1' position inside the R1 subunit active site, leading to a fast rate of appearance of λ at 326 nm. The single methylfuranone formed in all probability explains the second absorption value detected at 366 nm by chemically rearranging within the R1 active site, probably due to a different binding pose/site caused by the presence of the methyl group in the furanone species. This event would explain the complete disappearance of the

absorption value at 326 nm and formation of a new absorption band at 366 nm.

Acknowledgment. M.A.S.P. would like to thank the Fundação para a Ciência e a Tecnologia (FCT) for a Ph.D. grant.

Supporting Information Available: Energies of the reactants, transition states and products, provided by B3LYP (and BB1K) using the large model, are presented in the Supporting Information. The coordinates (together with charges, layers and connectivities) of each reactant, transition state and product are also presented in the Supporting Information. This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- (1) Lawrence, C. C.; Bennati, M.; Obias, H. V.; Bar, G.; Griffin, R. G.; Stubbe, J. High-field EPR detection of a disulfide radical anion in the reduction of cytidine 5'-diphosphate by the E441Q R1 mutant of Escherichia coli ribonucleotide reductase. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 8979–8984.
- (2) Robins, M. J.; Samano, V.; Zhang, W. J.; Balzarini, J.; Declercq, E.; Borchardt, R. T.; Lee, Y.; Yuan, C. S. Nucleic-Acid Related-Compounds 0.74. Synthesis and Biological-Activity of 2'(and 3')-Deoxy-2'(and 3')-Methylenenucleoside Analogs That Function as Mechanism-Based Inhibitors of S-Adenosyl-L-Homocysteine Hydrolase and or Ribonucleotide Reductase. *J. Med. Chem.* **1992**, *35*, 2283–2293.
- (3) Stubbe, J. A.; van der Donk, W. A. Ribonucleotide reductases: Radical enzymes with suicidal tendencies. *Chem. Biol.* **1995**, *2*, 793–801.
- (4) Cory, J. G. Ribonucleotide Reductase as a Chemotherapeutic Target. *Adv. Enzyme Regul.* **1988**, *27*, 437–455.
- (5) Nocentini, G. Ribonucleotide reductase inhibitors: New strategies for cancer chemotherapy. *Crit. Rev. Oncol./Hematol.* **1996**, *22*, 89–126.
- (6) Gerfen, G. J.; van der Donk, W. A.; Yu, G. X.; McCarthy, J. R.; Jarvi, E. T.; Matthews, D. P.; Farrar, C.; Griffin, R. G.; Stubbe, J. Characterization of a substrate-derived radical detected during the inactivation of ribonucleotide reductase from Escherichia coli by 2'-fluoromethylene-2'-deoxycytidine 5'-diphosphate. *J. Am. Chem. Soc.* **1998**, *120*, 3823–3835.
- (7) Zhou, B. B. S.; Elledge, S. J. The DNA damage response: putting checkpoints in perspective. *Nature* **2000**, *408*, 433–439.
- (8) Eklund, H.; Uhlin, U.; Farnegardh, M.; Logan, D. T.; Nordlund, P. Structure and function of the radical enzyme ribonucleotide reductase. *Prog. Biophys. Mol. Biol.* **2001**, *77*, 177–268.
- (9) Stubbe, J. A.; van der Donk, W. A. Protein radicals in enzyme catalysis (vol 98, pg 705, 1998). *Chem. Rev.* **1998**, *98*, 705–762.
- (10) Scott, C. P.; Kashlan, O. B.; Lear, J. D.; Cooperman, B. S. A quantitative model for allosteric control of purine reduction by murine ribonucleotide reductase. *Biochemistry* **2001**, *40*, 1651–1661.
- (11) Kashlan, O. B.; Scott, C. P.; Lear, J. D.; Cooperman, B. S. A comprehensive model for the allosteric regulation of mam-

- malian ribonucleotide reductase. Functional consequences of ATP- and dATP-induced oligomerization of the large subunit. *Biochemistry* **2002**, *41*, 462–474.
- (12) Barlow, T. Evidence for a New Ribonucleotide Reductase in Anaerobic Escherichia-Coli. *Biochem. Biophys. Res. Commun.* **1988**, *155*, 747–753.
- (13) Fontecave, M.; Eliasson, R.; Reichard, P. Oxygen-Sensitive Ribonucleoside Triphosphate Reductase Is Present in Anaerobic Escherichia-Coli. *Proc. Natl. Acad. Sci. U.S.A.* **1989**, *86*, 2147–2151.
- (14) van der Donk, W. A.; Yu, G. X.; Perez, L.; Sanchez, R. J.; Stubbe, J.; Samano, V.; Robins, M. J. Detection of a new substrate-derived radical during inactivation of ribonucleotide reductase from Escherichia coli by gemcitabine 5'-diphosphate. *Biochemistry* **1998**, *37*, 6419–6426.
- (15) Mohr, M.; Zipse, H. C-H bond activation in ribonucleotide reductases - Do short, strong hydrogen bonds play a role. *Chem.—Eur. J.* **1999**, *5*, 3046–3054.
- (16) Stubbe, J.; Yee, C. S.; Chang, M. C. Y.; Ge, J.; Nocera, D. G. Ribonucleotide reductase: Unnatural amino acids to probe proton coupled electron transfer. *Abstr. Pap. Am. Chem. Soc.* **2003**, *226*, 054-BIOL.
- (17) Stubbe, J. Radicals with a controlled lifestyle. *Chem. Commun.* **2003**, 2511–2513.
- (18) Zipse, H. The influence of hydrogen bonding interactions on the C-H bond activation step in class I ribonucleotide reductases. *Org. Biomol. Chem.* **2003**, *1*, 692–699.
- (19) Bennati, M.; Robblee, J. H.; Mugnaini, V.; Stubbe, J.; Freed, J. H.; Borbat, P. EPR distance measurements support a model for long-range radical initiation in E.coli ribonucleotide reductase. *J. Am. Chem. Soc.* **2005**, *127*, 15014–15015.
- (20) Strand, K. R.; Karlsen, S.; Kolberg, M.; Rohr, A. K.; Gorbitz, C. H.; Andersson, K. K. Crystal structural studies of changes in the native dinuclear iron center of ribonucleotide reductase protein R2 from mouse. *J. Biol. Chem.* **2004**, *279*, 46794–46801.
- (21) Baker, C. H.; Banzon, J.; Bollinger, J. M.; Stubbe, J.; Samano, V.; Robins, M. J.; Lippert, B.; Jarvi, E.; Resvick, R. 2'-Deoxy-2'-Methylenecytidine and 2'-Deoxy-2',2'-Difluorocytidine 5'-Diphosphates - Potent Mechanism-Based Inhibitors of Ribonucleotide Reductase. *J. Med. Chem.* **1991**, *34*, 1879–1884.
- (22) Takenuki, K.; Matsuda, A.; Ueda, T.; Sasaki, T.; Fujii, A.; Yamagami, K. Nucleosides and Nucleotides 0.83. Design, Synthesis, and Antineoplastic Activity of 2'-Deoxy-2'-Methylenecytidine. *J. Med. Chem.* **1988**, *31*, 1063–1064.
- (23) Masuda, N.; Matsui, K.; Yamamoto, N.; Nogami, T.; Nakagawa, K.; Negoro, S.; Takeda, K.; Takifuji, N.; Yamada, M.; Kudoh, S.; Okuda, T.; Nemoto, S.; Ogawa, K.; Myobudani, H.; Nihira, S.; Fukuoka, M. Phase I trial of oral 2'-deoxy-2'-methylidenecytidine: On a daily x 14-day schedule. *Clin. Cancer Res.* **2000**, *6*, 2288–2294.
- (24) Baker, C.; Banzon, J.; Bollinger, J.; Stubbe, J.; Samano, V.; Robins, M.; Lippert, B.; Jarvi, E.; Resvick, R. 2'-Deoxy-2'-Methylenecytidine and 2'-Deoxy-2',2'-Difluorocytidine 5'-Diphosphates - Potent Mechanism-Based Inhibitors of Ribonucleotide Reductase. *J. Med. Chem.* **1991**, *34*, 1879–1884.
- (25) Trucks, R.C., M. J. F., 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, and C. Pople, J. A. *Gaussian 03*; Gaussian, Inc.: Wallingford, CT, 2004.
- (26) Becke, A. D. Density-Functional Exchange-Energy Approximation with Correct Asymptotic-Behavior. *Phys. Rev. A* **1988**, *38*, 3098–3100.
- (27) Becke, A. D. Density-functional thermochemistry 0.4. A new dynamical correlation functional and implications for exact-exchange mixing. *J. Chem. Phys.* **1996**, *104*, 1040–1046.
- (28) Lee, C. T.; Yang, W. T.; Parr, R. G. Development of the Colle-Salvetti Correlation-Energy Formula into a Functional of the Electron-Density. *Phys. Rev. B* **1988**, *37*, 785–789.
- (29) Himo, F. Quantum chemical modeling of enzyme active sites and reaction mechanisms. *Theor. Chem. Acc.* **2006**, *116*, 232–240.
- (30) Cerqueira, N.; Fernandes, P. A.; Eriksson, L. A.; Ramos, M. J. Dehydration of ribonucleotides catalyzed by ribonucleotide reductase: The role of the enzyme. *Biophys. J.* **2006**, *90*, 2109–2119.
- (31) Cerqueira, N.; Fernandes, P. A.; Ramos, M. L. Understanding ribonucleotide reductase inactivation by gemcitabine. *Chem.—Eur. J.* **2007**, *13*, 8507–8515.
- (32) Leopoldini, M.; Marino, T.; Michelini, M. D.; Rivalta, I.; Russo, N.; Sicilia, E.; Toscano, M. The role of quantum chemistry in the elucidation of the elementary mechanisms of catalytic processes: from atoms, to surfaces, to enzymes. *Theor. Chem. Acc.* **2007**, *117*, 765–779.
- (33) Siegbahn, P. E. M.; Eriksson, L.; Himo, F.; Pavlov, M. Hydrogen atom transfer in ribonucleotide reductase (RNR). *J. Phys. Chem. B* **1998**, *102*, 10622–10629.
- (34) Eriksson, M.; Uhlin, U.; Ramaswamy, S.; Ekberg, M.; Regnstrom, K.; Sjoberg, B.; Eklund, H. Binding of allosteric effectors to ribonucleotide reductase protein R1: reduction of active-site cysteines promotes substrate binding. *Structure* **1997**, *5*, 1077–1092.
- (35) Case, D. A.; Darden, T. A.; Cheatham, T. E., III; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Merz, H. M.; Wang, B.; Pearlman, D. A.; Crowley, M.; Brozell, S.; Tsui, V.; Gohlke, H.; Mongan, J.; Hornak, V.; Cui, G.; Beroza, P.; Schafmeister, C.; Caldwell, J. W.; Ross, W. S.; Kollman, P. A. *AMBER 8*; University of California: San Francisco, CA, 2004.
- (36) Bayly, C. I.; Cieplak, P.; Cornell, W. D.; Kollman, P. A. A Well-Behaved Electrostatic Potential Based Method Using Charge Restraints for Deriving Atomic Charges - the Resp Model. *J. Phys. Chem.* **1993**, *97*, 10269–10280.
- (37) Basma, M.; Sundara, S.; Calgan, D.; Vernali, T.; Woods, R. J. Solvated ensemble averaging in the calculation of partial atomic charges. *J. Comput. Chem.* **2001**, *22*, 1125–1137.

- (38) Kirschner, K. N.; Woods, R. J. Solvent interactions determine carbohydrate conformation. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 10541–10545.
- (39) Kirschner, K. N.; Woods, R. J. Quantum mechanical study of the nonbonded forces in water-methanol complexes. *J. Phys. Chem. A* **2001**, *105*, 4150–4155.
- (40) Dapprich, S.; Komaromi, I.; Byun, K. S.; Morokuma, K.; Frisch, M. J. A new ONIOM implementation in Gaussian98. Part I. The calculation of energies, gradients, vibrational frequencies and electric field derivatives. *J. Mol. Struct.: THEOCHEM* **1999**, *461*, 1–21.
- (41) Maseras, F.; Morokuma, K. Imom - a New Integrated Ab-Initio Plus Molecular Mechanics Geometry Optimization Scheme of Equilibrium Structures and Transition-States. *J. Comput. Chem.* **1995**, *16*, 1170–1179.
- (42) Cerqueira, N.; Fernandes, P.; Eriksson, L.; Ramos, M. Dehydration of ribonucleotides catalyzed by ribonucleotide reductase: The role of the enzyme. *Biophys. J.* **2006**, *90*, 2109–2119.
- (43) Cioslowski, J. A New Population Analysis Based on Atomic Polar Tensors. *J. Am. Chem. Soc.* **1989**, *111*, 8333–8336.
- (44) Fernandes, P. A.; Ramos, M. J. Theoretical studies on the mechanism of inhibition of ribonucleotide reductase by (E)-2'-fluoromethylene-2'-deoxycytidine-5'-diphosphate. *J. Am. Chem. Soc.* **2003**, *125*, 6311–6322.
- (45) Sousa, S. F.; Fernandes, P. A.; Ramos, M. J. General performance of density functionals. *J. Phys. Chem. A* **2007**, *111*, 10439–10452.
- (46) Cerqueira, N.; Fernandes, P. A.; Ramos, M. J. Enzyme ribonucleotide reductase: Unraveling an enigmatic paradigm of enzyme inhibition by furanone derivatives. *J. Phys. Chem. B* **2006**, *110*, 21272–21281.
- (47) Fernandes, P. A.; Eriksson, L. A.; Ramos, M. J. The reduction of ribonucleotides catalyzed by the enzyme ribonucleotide reductase. *Theor. Chem. Acc.* **2002**, *108*, 352–364.

CT1002175

Gold(I)-Catalyzed Hydration of 1,2-Diphenylacetylene: Computational Insights

Gloria Mazzone, Nino Russo, and Emilia Sicilia*

Dipartimento di Chimica, Università della Calabria, I-87030, Arcavacata di Rende, Italy

Received May 14, 2010

Abstract: A DFT investigation of 1,2-diphenylacetylene hydration mediated by the $[(\text{Ph}_3\text{P})\text{Au}]^+$ complex has been carried out to shed light on the mechanistic details of such process with the support of the experimental observations and mechanistic proposal. Computational analysis proves that the first inner-sphere attack of water occurs with gold acting as a proton shuttle to transfer the migrating hydrogen in *cis* position with respect to OH group. From the formed *E* isomer of the enol the *Z* one could be formed by rotation around the C–C bond. The addition of the second water molecule to give the ketone final product occurs favorably with the assistance of the catalyst and involves coordination of water followed by a second hydrogen shift from oxygen to carbon. If the *E* isomer is involved, gold directly participates in the reaction assisting the hydrogen transfer, whereas if the product is obtained starting from the *Z* isomer, gold is not directly involved. The elimination of a water molecule and the release of the catalyst close the catalytic cycle. Calculations show that the intervention of a third water molecule lowers the energy barrier for the elimination of water and formation of the π carbonyl bond.

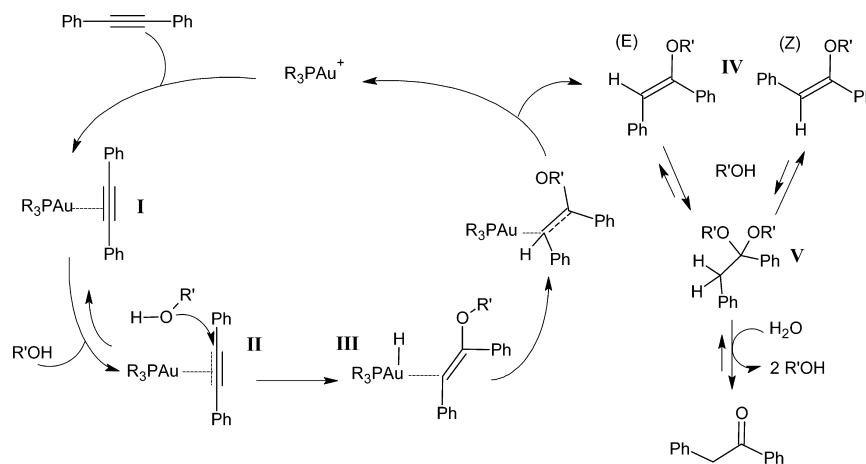
1. Introduction

The hydration of unsaturated carbon compounds is one of the most environmentally benign and economically attractive route to form a carbon–oxygen bond. Unactivated alkynes are an abundant hydrocarbon resource and hydration represents an excellent means of functionalization, which has been extensively studied.^{1–5}

It has long been known that water can be added quite easily to activated electron-rich alkynes in the presence of an acid catalyst.^{6,7} The reaction of simple alkynes, instead, needs cocatalysts, typically mercury(II) salts, to enhance the reactivity.^{8–14} Such catalysts have been extensively used in high-scale industrial processes¹⁵ until the discovery of the toxicity of mercury salts. Moreover, under the reaction conditions the mercury(II) is quickly reduced to catalytically inactive metallic mercury. To avoid the use of toxic mercury salts there has been much effort aimed at the development of alternative synthetic strategies for the hydration of alkynes based on the use of metal catalytic systems,¹⁶ but neither of them has shown activity comparable to that of those mercury

salt containing systems. Among the alternative systems, gold compounds have gradually taken a prominent place, beneficially replacing mercury salts. A notable landmark in the application of gold in this field is the report, after 90 years, of Teles and co-workers, in which the addition of methanol to alkynes has been reported to be efficiently catalyzed by coordinatively unsaturated cationic gold(I) species, of the type R_3PAu^+ , generated in situ by the protonolysis of R_3PAuCH_3 with an acid and release of CH_4 .^{17,18} A remarkable result of Teles work stems from theoretical calculations for the reaction pathway that show how gold(I) directs the alcohol nucleophile to the side of the coordinated metal. This behavior is different from the usual attack of the nucleophile from the side opposite to the coordinated metal that is assumed for other electrophilic metals. In 2002, Tanaka et al. using similar reagents achieved the efficient hydration of a wide range of alkynes in aqueous methanol.¹⁸ Good yields have been reported also by Laguna¹⁹ and others,²⁰ especially when gold(I) complexes were combined with strong acids under heating. Very recently, Leyva and Corma have carried out the efficient hydration of a wide range of alkynes to the corresponding ketones by using gold(I)–phosphine complexes as catalyst with no acidic cocatalyst.²¹

* To whom correspondence should be addressed. E-mail: siciliae@unical.it.

Scheme 1. Proposed Catalytic Cycle for the Alkynes Hydration in Presence of AuPR_3^+ Catalysts

Catalysts like R_3PAuX ($\text{X} =$ trifluoromethanesulfonate $^- \text{OTf}$ or other weakly coordinating counteranions) have been formed in situ by treatment of the corresponding chloride complex, R_3PAuCl , with a silver salt. The soft noncoordinating anions make the catalytic center acidic enough to bypass the use of additives. To shed light on the proposed mechanistic hypotheses,^{17–19} even contradictory, for this reaction the authors have carried out a series of kinetic experiments using H_2O , MeOH , or both as nucleophiles; 1,2-diphenylacetylene as substrate; and AuPR_3X ($\text{PR}_3 = \text{PPh}_3$, SPhos , P^tBu_3 ; $\text{X} = \text{Cl}$, OTf , NTf_2) as catalyst in THF solvent.

The mechanism proposed by the authors on the basis of experimental findings involves (Scheme 1) coordination of the triple bond to Au(I) complex with formation of the $\text{Au}-\pi$ -alkyne complex **I** and subsequent attack of the $\text{R}'\text{OH}$ ($\text{R}' = \text{H}$, CH_3) nucleophile. $\text{R}'\text{OH}$ would coordinate to gold, leading to formation of the intermediate **II**, before its addition to the triple bond (**III**) to afford, by protodeauration, the *E* isomer of the enol ether product **IV** together with the restored catalyst. Formation of the *Z* from the *E* isomer can occur by rotation about the $\text{C}-\text{C}$ double bond¹⁷ or in the next step of the overall process, that is, addition of a second molecule of $\text{R}'\text{OH}$ to form the corresponding ketal **V**. Elimination of a $\text{R}'\text{H}$ molecule from the **V** would explain formation of both *Z* and *E* isomers, whereas subsequent reaction of ketal **V** with H_2O leads to the formation of the ketone final product. The authors suggest that the gold catalyst could come into play also in these last steps of the overall hydration process.

As the first step of a more comprehensive project, we have investigated here, with the aid of density functional theory (DFT) calculations, the detailed mechanism of the whole of 1,2-diphenylacetylene hydration process catalyzed by the gold(I) $[(\text{Ph}_3\text{P})\text{Au}]^+$ complex using pure water as nucleophile with the support of the experimental observations and mechanistic proposal. The water is able to attack the alkyne from the same side of the coordinated catalyst (inner-sphere mechanism) or from the opposite side (outer-sphere mechanism). Even if not envisaged by Teles and Leyva and Corma, the outer-sphere attack of H_2O to the alkyne has been also explored.

The theoretical insights presented in this work are expected to be helpful in understanding such kind of gold-catalyzed

hydration reactions and provide helpful information for chemists on similar processes.

2. Computational Details

All molecular geometries have been optimized at the Becke3-LYP (B3LYP) level of density functional theory.^{22,23} Numerous theoretical studies of Au-catalyzed reactions at the B3LYP level have been reported in the literature, which confirm that such exchange-correlation functional is quite suitable to investigate Au-catalyzed reactions.^{24–27} However, to check the effect of the exchange-correlation functional, additional calculations employing the B97-1²⁸ functional have been carried out. B97-1 functional, indeed, has been shown by Zhao and Truhlar to provide energetics of comparable quality to MP2 results for weakly interacting systems.²⁹

For both B3LYP and B97-1 functionals, frequency calculations at the same level of theory have been also performed to identify all stationary points as minima (zero imaginary frequencies) or transition states (one imaginary frequency).

The transition states involved have been checked by IRC (intrinsic reaction coordinate) analysis.^{30,31} For Au, the relativistic compact Stuttgart/Dresden effective core potential³² has been used in conjunction with its split valence basis set. In order to reduce the computational cost of stationary points, optimization standard 6-31G basis set of Pople for carbon and hydrogen atoms of phenyl rings has been used together with the 6-31G** for the rest of the atoms. Final energies have been calculated by performing single-point calculations on the optimized geometries at the same level of theory and employing 6-311+G** standard basis sets for H, C, O and P atoms. All the calculations have been performed with the Gaussian 03 software package.³³

The impact of solvation effects on the energy profiles has been estimated by using the conductor polarizable continuum model (CPCM)³⁴ as implemented in Gaussian 03. The UAHF set of radii has been used to build up the cavity. Since preliminary calculations have clearly shown geometry relaxation effects to be not significant, the solvation Gibbs free energies have been calculated in implicit THF ($\epsilon = 7.58$), the solvent medium in the experiments, with the above

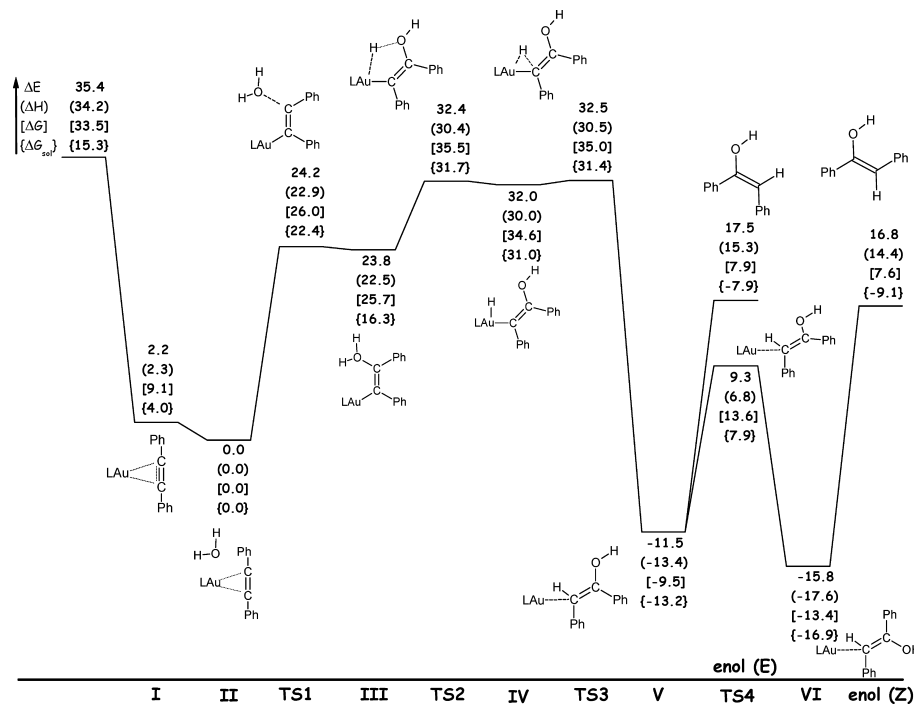


Figure 1. Calculated B3LYP PES for the addition of the first water molecule to 1,2-diphenylacetylene catalyzed by the gold(I) $[(\text{Ph}_3\text{P})\text{Au}]^+$ complex. Relative ZPE-corrected electronic energies, enthalpies, and Gibbs free energies at 298.15 K together with free energy changes in THF are reported. Relative energies are in kcal/mol.

method performing single-point calculations on all stationary points structures obtained from vacuum calculations at the B3LYP level. Other authors also have observed the negligible influence of solute geometry relaxation in solution.^{35–37} Reaction Gibbs free energies in solution, ΔG_{sol} , have been calculated for each process as the sum of two contributions: a gas-phase reaction free energy, ΔG_{gas} , and a solvation reaction free energy term calculated with the continuum approach, ΔG_{solv} . Single point energies have been also computed at the MP2 level through the RI-MP2³⁸ approach as implemented in the TURBOMOLE program package³⁹ (version 5.10) in conjunction with the same relativistic Stuttgart/Dresden pseudopotential for gold and the standard internally stored TZVP basis set^{40,41} for the rest of the atoms (see Figures S1 and S2 of the Supporting Information).

3. Results and Discussion

3.1. First H₂O Molecule Addition. B3LYP and B97-1 calculated energy profiles for the reaction pathway corresponding to the addition of the first H₂O molecule are shown in Figure 1 and Figure S3 of Supporting Information, respectively. Relative ZPE corrected electronic energies (ΔE), enthalpies (ΔH), and Gibbs free energies (ΔG) at 298 K, as well as free energies in THF (ΔG_{sol}) are provided. Unless otherwise noted, in what follows, the discussed energies are B3LYP-relative ZPE-corrected electronic energies calculated with respect to the intermediate named **II**.

Structures of reactants, intermediates, transition states, and products of the reactions are schematically depicted in the same figure. Fully optimized B3LYP and B97-1 structures of stationary points can be found in Figures S6 and S7 of the Supporting Information, respectively.

From the energy profiles in Figure 1, it is evident that the first step along the pathway for the catalytic addition of water to the alkyne involves a preliminary intermediate, **I**, where the C–C triple bond interacts with the gold atom. Intermediate **I** is formed without any barrier and is 33.2 kcal/mol lower in energy than $[(\text{Ph}_3\text{P})\text{Au}]^+ + \text{C}_2\text{Ph}_2 + \text{H}_2\text{O}$ reactants. Several examples^{42–46} of such a π -coordination to gold(I) have been theoretically studied to evaluate the extent of the π -to-metal σ -donation and metal-to- π^* back-donation, and the main conclusion is that Au(I) cation cannot significantly participate to a Dewar–Chatt–Duncanson-type bonding,^{47,48} as antibonding orbitals are too high in energy for significant back-bonding to occur. Here the C1–C2 bond has lost a little of its triple bond character. Indeed, the C–C length increases from 1.205 Å, calculated for the free alkyne, to 1.239 Å, and the deviation from linearity is measured by the value of 169.5° of the C1–C2–C4 angle. With these optimized geometries in hand, the natural bond order (NBO)^{49,50} analysis have been used to investigate the nature of the metal–alkyne bond. Second-order perturbative analysis has revealed that π -to-metal σ -donation largely dominates (105.8 kcal/mol) over metal-to- π^* back-donation (14.1 kcal/mol). In addition to NBO analysis also the frontier molecular orbitals have been examined (see Figure 2). As expected, the LUMO contains the π^* –metal interaction. Furthermore, results from the calculation of the energies of the frontier orbitals of the two species $[(\text{Ph}_3\text{P})\text{Au}]^+$ and C_2Ph_2 show that the HOMO of diphenylacetylene is only 0.3 eV higher in energy than the LUMO of the gold(I) complex. The difference in energy, instead, between the HOMO of $[(\text{Ph}_3\text{P})\text{Au}]^+$ and the LUMO of C_2Ph_2 is very large, that is, 9.4 eV. The lack of back-bonding from Au(I) enhances the electrophilicity of the ligand and favors the attack of the

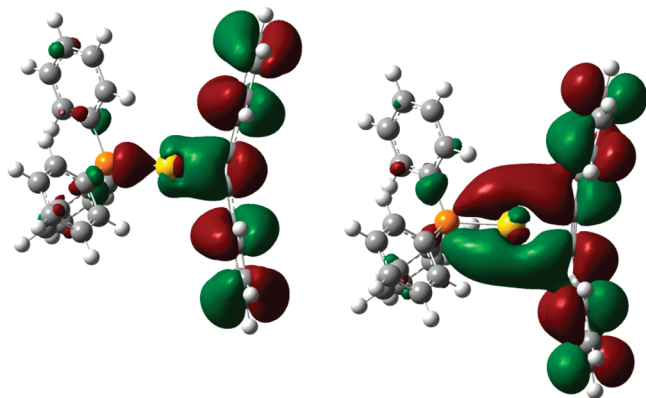


Figure 2. Structure of the highest occupied and lowest unoccupied molecular orbitals of the $[(\text{Ph}_3\text{P})\text{Au-alkyne}]^+$ complex.

nucleophile. As a first step, the intermediate **II** is formed in which water is loosely coordinated to the metal. This precoordination is computed to be exothermic by 2.2 kcal/mol and introduces some changes into the structure of the complex. The lengths of the two Au–C bonds, that are 2.332 Å, become different (Au–C1 = 2.356 Å and Au–C2 = 2.306 Å). The angle between phosphorus, gold, and the next C2 carbon of the triple bond changes from 164.3° to 156.9°. The Au–O bond is 3.039 Å and the P–Au–O angle is 82.9°.

The next step involves formation of intermediate **III** by coordination between the O-atom of water and one carbon atom of the diphenylacetylene concomitant with a movement of Au⁺ to the other carbon. The calculated activation energy for rearrangement of **II** to **III** is 24.2 kcal/mol, and the intermediate **III** is only slightly more stable than the transition state leading to it. The further reaction toward formation of the *E* isomer of the enol coordinated to the Au(I)-complex involves an Au⁺-assisted hydrogen migration from oxygen, via gold, to the terminal carbon to finally afford the product complex **V** with a relative energy of –11.5 kcal/mol. The first step, that is, migration to Au⁺, has a computed activation energy of 8.6 kcal/mol, and the intermediate **IV** is formed that lies 32 kcal/mol higher in energy with respect to reference adduct **II**. The next hydrogen shift from gold to the position cis to the OH group corresponds to a very low activation energy of about 0.5 kcal/mol. Even if the very small differences in energy among the **TS2** and **TS3** transition states and **IV** intermediate could suggest the involvement of a concerted rearrangement, all the attempts to intercept a transition state directly connecting **III** and **V** intermediates have been unsuccessful. MP2 calculations (see Figure S1 of the Supporting Information) confirm this result.

Our calculations show that the *Z* isomer formation occurs exclusively by rotation around the C–C double bond of the complex **V**, since no pathway exists for the direct transfer of the hydrogen atom from oxygen to the position trans to the OH group. The transition state structure for the catalyst-assisted *E* to *Z* isomerization is located at a relative energy of 9.3 kcal/mol and corresponds to an activation energy of 20.8 kcal/mol, which is significantly lower than the barrier to rotation around a double bond in the absence of catalyst, both experimentally and theoretically estimated to be approximately 65 kcal/mol.^{51,52}

The *Z* isomer coordinated to the gold complex is stabilized by 4.3 kcal/mol with respect to the *E* one. As pointed out in recent literature,^{53,54} even if gold–alkene π -complexes are known for a long time, the number of such structurally characterized complexes is quite limited. From our computational analysis results, for both isomers the binding of Au⁺ to the double bond is very asymmetric in that the gold atom is only attached to the C2 terminal carbon atom. The C1–C2 bond length is 1.402 Å for the *E* isomer and 1.398 Å for the *Z* isomer. The C1–C2–Au angle amounts to 90.7° and 89.7° in *E* and *Z* isomers, respectively. NBO analysis reveals that even if π -to-metal σ -donation plays a role, electrostatic interactions are the primary origins of the bond. Both NBO and ESP^{55,56} charge analyses show that the positive charge is significantly redistributed on the ligand and the gold atom carries a charge not larger than 0.29. An attractive interaction, therefore, exists with the C2 atom of the alkene that carries a negative charge, whereas the C1 atom is positively charged due to the bond with the O atom. More details can be found in Figure S8 of the Supporting Information. The catalytic cycle could be, then, closed, either by the *Z* isomer of the enol release that is calculated to be endothermic by 32.6 kcal/mol or by the loss of the *E* isomer that requires 29 kcal/mol to occur.

To close this section concerning the addition of the first water molecule, we report the results obtained by the investigation of the outer-sphere mechanism. Figure 3 presents the B3LYP energy profile for the water attack from the opposite side with respect of the coordinated catalyst, whereas selected geometrical parameters of stationary points are shown in Figure S9 of Supporting Information.

Once the preliminary intermediate **I** is formed, further stabilization is due to the loose precoordination of the water molecule that weakly interacts with one of the triple bond carbon atoms. This step, calculated to be the exothermic by 3.1 kcal/mol, is followed by the simultaneous formation of the C1–O bond and the cleavage of one O–H bond, accompanied by the transfer of the hydrogen atom to the C2 atom. This rearrangement, corresponding to the formation of the **TS1'** transition state, has a very high energetic cost of 60.6 kcal/mol and leads to the formation of the *E* isomer of the enol coordinated to catalyst. It is worth noting that the **TS1'** transition state rearrangement involves the coordination of the gold catalyst to the carbon atom of the C1–C2 alkene bond being replaced by the interaction of the gold atom with one of the phenyl rings of the alkyne. Therefore, to obtain the final complex **V** it is necessary to overcome a low energy barrier, of 4.4 kcal/mol, corresponding to the transition state, **TS2'**, which allows gold to restore the initial coordination to the C–C bond of the alkene. On the basis of the very high energy barrier calculated along the outer-sphere nucleophilic attack pathway, these results strongly support an inner-sphere mechanism.

3.2. Second H₂O Molecule Addition. As underlined by the authors,²¹ the catalyst could play also a role in the next steps of the whole process that involves the addition of a second water molecule. To evaluate which is the most favorable pathway, we have calculated the energy profiles for the reaction occurring both in the absence of and assisted

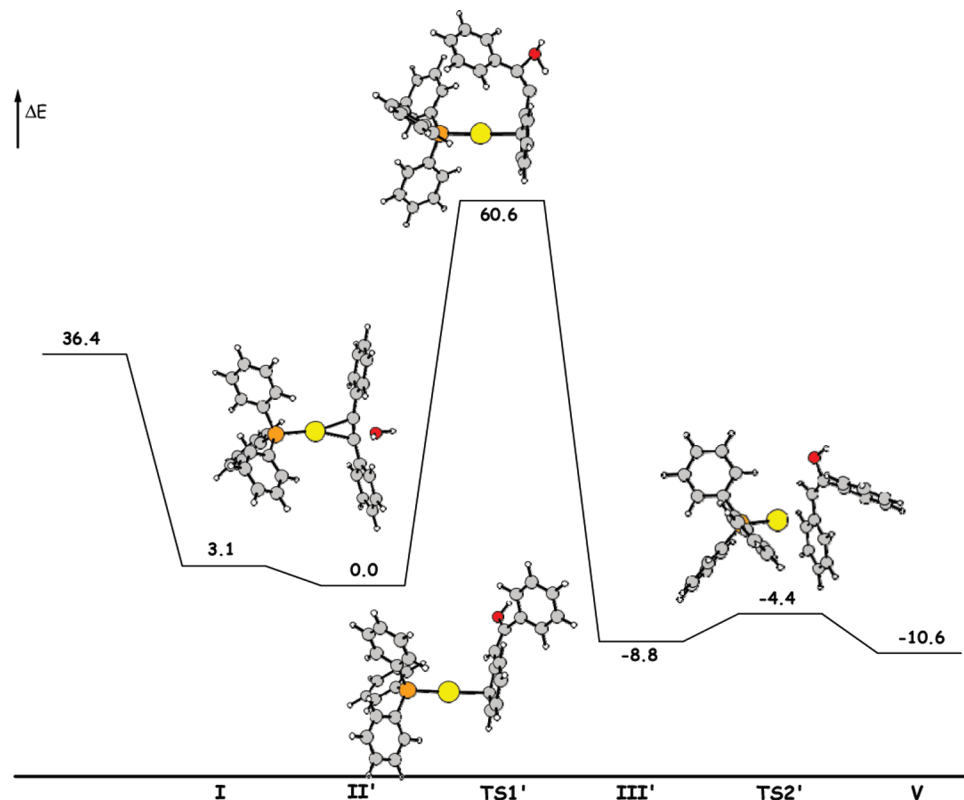


Figure 3. Calculated B3LYP energy profile for the outer-sphere attack of the first water molecule to 1,2-diphenylacetylene catalyzed by the gold(I) $[(\text{Ph}_3\text{P})\text{Au}]^+$ complex. Relative ZPE-corrected electronic energies, in kcal/mol, are reported.

by the gold catalyst. B3LYP- and B97-1-calculated energy profiles for the reaction pathway corresponding to the addition of the second water molecule are shown in Figure 4 and Figure S4 of Supporting Information, respectively. We are going to comment only on B3LYP-relative ZPE-corrected electronic energies (ΔE), calculated with respect to the **II** intermediate, even if also enthalpies (ΔH) and Gibbs free energies (ΔG) at 298 K as well as free energies in THF (ΔG_{sol}) have been calculated and reported. In the same Figure 4, the B3LYP energy profile for the direct addition of water in the absence of the gold catalyst is also shown (see Figure S10 of the Supporting Information for more details). Fully optimized B3LYP and B97-1 structures of stationary points intercepted along the catalyst-assisted reaction pathway can be found in Figures S11 and S12 of Supporting Information, respectively.

Along the catalyst-assisted pathway both **V** and **VI** adducts, that is, *E* and *Z* isomers coordinated to the $[(\text{Ph}_3\text{P})\text{Au}]^+$ complex, can enter the subsequent catalytic cycle for the addition of a second water molecule. Indeed, the possibility that the isomerization from *E* to *Z* isomers occurs depends on how rapidly the final ketone product is formed. For that reason, we have investigated the pathway for the addition of a second H_2O molecule starting from both **V** (solid line in Figure 4) and **VI** (dashed line in Figure 4) intermediates.

The B3LYP-calculated energy profiles in Figure 4 show that the first step of the process is the loose coordination of H_2O that is calculated to be exothermic by 19.8 kcal/mol and 25.7 with respect to **II** intermediate along *E* and *Z* enol isomers pathways, respectively. An examination of Figure 4 shows that along the path involving the *E* isomer the reaction proceeds by a 1,3-hydrogen migration with gold

acting as a proton shuttle. Oxygen coordinates to C2 and one of the hydrogen atoms is transferred to Au to form the **VIIIa** intermediate that lies 28.5 kcal/mol above the energy of **II** intermediate. The barrier for this rearrangement (**TS5a**) is 48.8 kcal/mol, whereas the barrier for the next hydrogen shift from Au to the carbon atom (**TS6a**) to form the corresponding *gem*-diol **IX** is very low and amounts to 1.0 kcal/mol. Although on the basis of the flatness of the PES in this region one can suspect that the reaction occurs in one step, B97-1 analysis, as well as MP2 computations (see Figure S2 of Supporting Information), confirm the existence of two distinct transition states and of the minimum between them. Along the *Z* isomer pathway, instead, the intermediate **IX** formation takes place directly (**TS5b**), since the hydrogen atom is transferred from oxygen to carbon in one step and the barrier that is necessary to overcome is 55.2 kcal/mol. Therefore, the calculated barrier for the *E* to *Z* isomerization (first H_2O molecule addition) is lower than both the barrier that must be overcome to generate from the *E* isomer the **IX** intermediate and the reverse barrier to go from **IX** to the *Z* isomer by elimination of a water molecule (45.9 kcal/mol). Hence, it should be concluded that formation of the *Z* isomer is likely to occur by direct isomerization from the *E* isomer. The final step of the reaction, that is, formation of the ketone product from intermediate **IX**, can occur both by direct elimination of a water molecule from the two OH groups and by the involvement of a third molecule of water. The transition states for the both pathways have been intercepted. In Figure 5 are reported the B3LYP energetics of the process involving an additional water molecule assisting the elimination reaction, leading to formation of the ketone coordinated

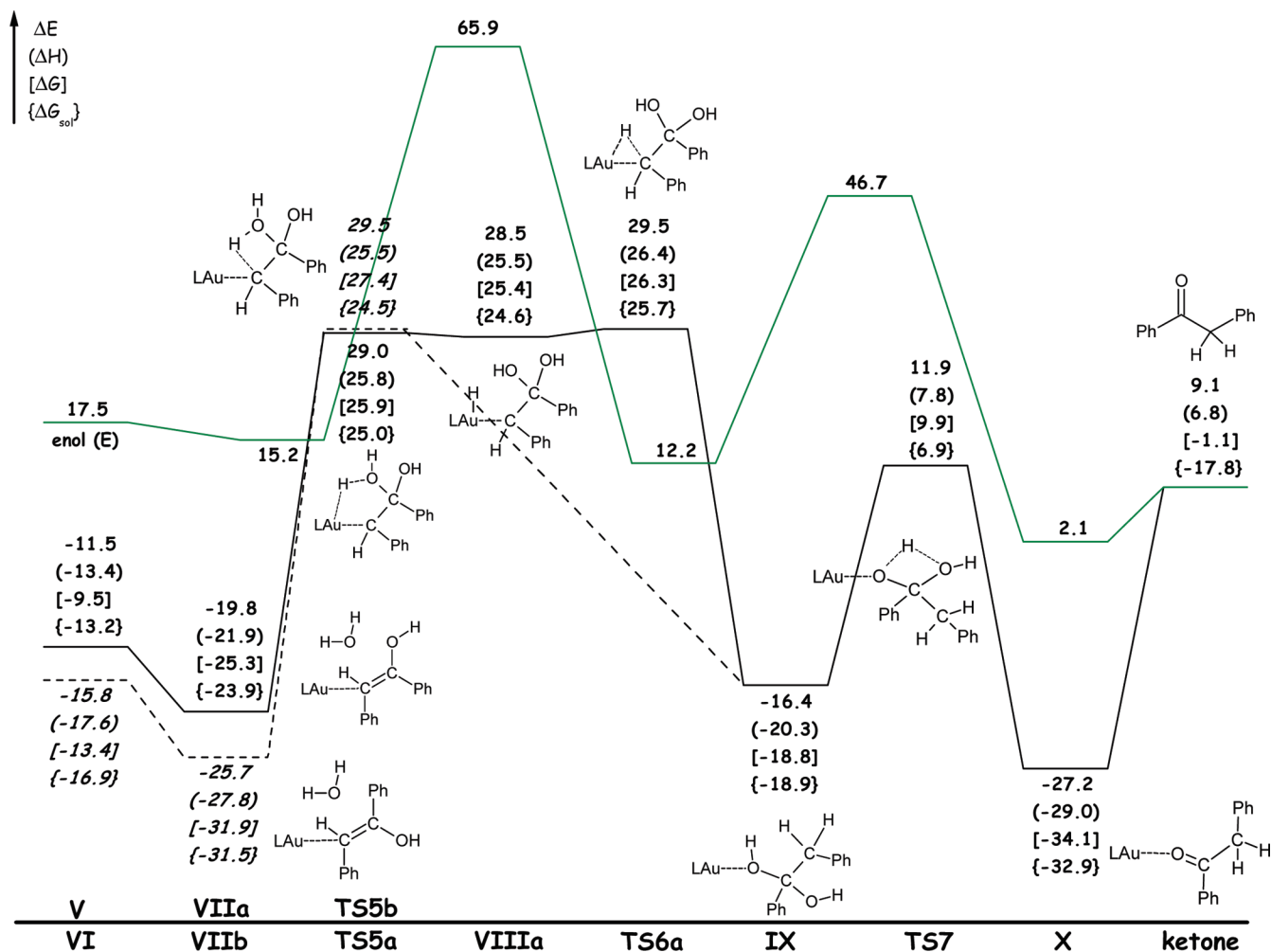


Figure 4. Calculated B3LYP energy profiles for the addition of the second water molecule to enol *E* (solid line) and *Z* (dashed line) isomers coordinated to the $[(\text{Ph}_3\text{P})\text{Au}]^+$ complex. Relative ZPE-corrected electronic energies (ΔE), enthalpies (ΔH), and Gibbs free energies (ΔG) at 298.15 K together with free energy changes in THF (ΔG_{sol}) are reported. The energetics of the uncatalyzed pathways is also reported (green line). Relative energies are in kcal/mol.

yet to the catalyst. The same energetics explored with the B97-1 functional can be found in Figure S5 of the Supporting Information. B3LYP and B97-1 geometrical structures of involved stationary points are included in Figures S11 and S12, respectively.

As can be clearly inferred from the comparison between Figures 4 and 5, the process is calculated to be more favorable if the participation of a third water molecule is taken into consideration. The elimination of a water molecule directly from the intermediate IX occurs through the TS7 transition state, whose formation is endothermic by 11.9 kcal/mol relative to the reference energy of the intermediate II and the corresponding activation barrier amounts to 28.3 kcal/mol. As a result of a hydrogen shift from one oxygen to the other, a water molecule is eliminated and the carbonyl bond is formed (minimum X). The TS7' transition state, instead, lies at 7.8 kcal/mol below the reference energy of II and corresponds to a barrier of 8.6 kcal/mol. The calculated imaginary frequency is associated with the movement of the third water molecule that approaches the complex and induces formation of the product by abstracting a hydrogen atom from one of the two OH groups and simultaneously transferring a hydrogen atom to the other OH. The formation

of the adduct X' with two directly interacting H₂O molecules is calculated to be slightly more exothermic than the formation of the X adduct with one water molecule, 30.7 kcal/mol versus 27.2 kcal/mol. To release the product and to regenerate the catalyst require overcoming an energy barrier of 36.3 and 39.8 kcal/mol from intermediate X and X', respectively. The whole catalytic process is endothermic by 9.1 kcal/mol if intermediate II is assumed to be the reference, whereas with respect to $[(\text{Ph}_3\text{P})\text{Au}]^+ + \text{C}_2\text{Ph}_2 + \text{H}_2\text{O}$ starting reactants the whole process is exothermic by 26.3 kcal/mol. Comparison in Figure 4 of the catalyzed and uncatalyzed pathways clearly shows that the intervention of the catalyst also in the second part of the process is crucial to lower the energy barriers and to act in such a way that all the stationary points along the reaction path lie below the energy of the reactants asymptote.

4. Conclusions

The investigation of the whole hydration process of 1,2-diphenylacetylene catalyzed by the gold(I) $[(\text{Ph}_3\text{P})\text{Au}]^+$ complex by addition of water to form the corresponding ketone has been carried out and the mechanistic hypoth-

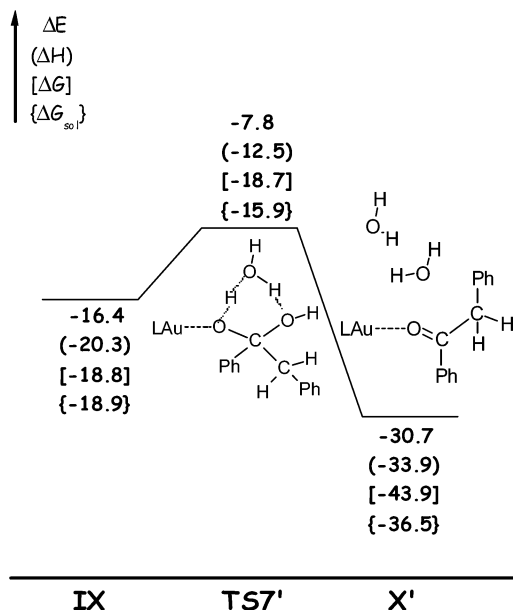


Figure 5. Calculated B3LYP energy profile for the water elimination step, leading to formation of the ketone coordinated to the catalyst, assisted by a third H_2O molecule. Relative ZPE-corrected electronic energies (ΔE), enthalpies (ΔH), and Gibbs free energies (ΔG) at 298.15 K together with free energy changes in THF (ΔG_{sol}) are reported. Energies are in kcal/mol and relative to intermediate II.

eses existing in the literature have been explored. In total, we can summarize our findings as follows. Calculations confirm that the first molecule addition occurs with gold acting as a proton shuttle to transfer the migrating hydrogen in cis position with respect to OH group. From the formed *E* isomer of the enol coordinated to the catalyst the *Z* one could be formed by rotation around the C–C bond. The addition of the second water molecule to give the ketone final product occurs favorably with the support of the catalyst and involves a second hydrogen shift from oxygen to carbon. If the *E* isomer is involved, gold directly participates in the reaction, assisting the hydrogen transfer, whereas if the product is obtained starting from the *Z* isomer, gold is not directly involved. The intervention of a third water molecule lowers the energy barrier for the final elimination of a water molecule and formation of the π carbonyl bond. The theoretical insights presented in this work are expected to help us understand this gold-catalyzed hydration reaction as well as similar processes. In this perspective, work is in progress to carry out analogous calculations for the analysis of the mechanistic details of the 1,2-diphenylacetylene hydration with methanol as nucleophile.

Acknowledgment. This research has been supported by Università della Calabria. Financial support from MIUR-PRIN (2007-prot. 20077EZFR4_002) is gratefully acknowledged.

Supporting Information Available: MP2 energy profiles, selected geometrical parameters for all B3LYP- and B97-1-calculated structures of stationary points, charge distribution analysis for the $[(Ph_3P)Au-alkene]^+$ complexes, and selected geometrical parameters of B3LYP-calculated

stationary points intercepted along the pathways for the outer-sphere attack of water and uncatalyzed second water molecule addition. This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- (1) Hudrlik, P. F.; Hudrlik, A. M. In *The Chemistry of the Carbon-Carbon Triple Bond*; Patai, S., Ed.; Wiley: New York, 1978; Part 1, p 199.
- (2) Schmid, G. H. In *The Chemistry of the Carbon-Carbon Triple Bond*; Patai, S., Ed.; Wiley: New York, 1978; Part 1, p 275.
- (3) March, J. *Advanced Organic Chemistry*; Wiley: New York, 1992; p 762.
- (4) Larock, R. C.; Leong, W. W. In *Comprehensive Organic Synthesis*; Trost, B. M., Fleming, I., Semmelhack, M. F., Eds.; Pergamon: Oxford, 1991; Vol. 4, p 269.
- (5) Kozhevnikov, I. V. *Chem. Rev.* **1998**, *98*, 171–198.
- (6) Drenth, W.; Hogeveen, H. *Recl. Trav. Chim. Pays-Bas* **1960**, *79*, 1002.
- (7) Allen, A. D.; Chiang, Y.; Kresge, A. J.; Tidwell, T. T. *J. Org. Chem.* **1982**, *47*, 775–779, and references therein.
- (8) Hinton, H. D.; Niewland, J. A. *J. Am. Chem. Soc.* **1930**, *52*, 2892–2896.
- (9) Killian, D. B.; Hennion, G. F.; Niewland, J. A. *J. Am. Chem. Soc.* **1934**, *56*, 1384–1385.
- (10) Hennion, G. F.; Niewland, J. A. *J. Am. Chem. Soc.* **1935**, *57*, 2006–2007.
- (11) Killian, D. B.; Hennion, G. F.; Niewland, J. A. *J. Am. Chem. Soc.* **1936**, *58*, 1658–1659.
- (12) Killian, D. B.; Hennion, G. F.; Niewland, J. A. *J. Am. Chem. Soc.* **1936**, *58*, 80–81.
- (13) Hennion, G. F.; Murray, W. S. *J. Am. Chem. Soc.* **1942**, *64*, 1220–1222.
- (14) Bassetti, M.; Floris, B. *J. Chem. Soc. Perkin Trans. 2* **1988**, 227–233.
- (15) Acetaldehyde. *Ullmann's Encyclopedia of Industrial Chemistry*, 7th ed.; Wiley-VCH: Weinheim, 2006.
- (16) Hintermann, L.; Labonne, A. *Synthesis* **2007**, *8*, 1121–1150.
- (17) Teles, J. H.; Brode, S.; Chabanas, M. *Angew. Chem., Int. Ed.* **1998**, *37*, 1415–1418.
- (18) Mizushima, E.; Sato, K.; Hayashi, T.; Tanaka, M. *Angew. Chem., Int. Ed.* **2002**, *41*, 4563–4565.
- (19) Casado, R.; Contel, M.; Laguna, M.; Romero, P.; Sanz, S. *J. Am. Chem. Soc.* **2003**, *125*, 11925–11935.
- (20) Roembke, P.; Schmidbaur, H.; Cronje, S.; Raubenheimer, H. *J. Mol. Catal. A* **2004**, *212*, 35–42.
- (21) Leyva, A.; Corma, A. *J. Org. Chem.* **2009**, *74*, 2067–2074.
- (22) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648–5652.
- (23) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. *J. Phys. Chem.* **1994**, *98*, 11623–11627.
- (24) Roithová, J.; Hrušák, J.; Schröder, D.; Schwarz, H. *Inorg. Chim. Acta* **2005**, *358*, 4287–4292.
- (25) Nieto-Oberhuber, C.; López, S.; Jiménez-Núñez, E.; Echavarrén, A. M. *Chem.—Eur. J.* **2006**, *12*, 5916–5923.

- (26) Shi, F.-Q.; Li, X.; Xia, Y.; Zhang, L.; Yu, Z.-X. *J. Am. Chem. Soc.* **2007**, *129*, 15503–15512.
- (27) Kovács, G.; Ujaque, G.; Lledós, A. *J. Am. Chem. Soc.* **2008**, *130*, 853–864.
- (28) Hamprecht, F. A.; Cohen, A. J.; Tozer, D. J.; Handy, N. C. *J. Chem. Phys.* **1998**, *109*, 6264–6271.
- (29) Zhao, Y.; Truhlar, D. G. *J. Chem. Theory Comput.* **2005**, *1*, 415–432.
- (30) Fukui, K. *J. Phys. Chem.* **1970**, *74*, 4161–4163.
- (31) Gonzalez, C.; Schlegel, H. B. *J. Chem. Phys.* **1989**, *90*, 2154–2161.
- (32) Andrae, D.; Häussermann, U.; Dolg, M.; Stoll, H.; Preuss, H. *Theor. Chim. Acta* **1990**, *77*, 123–141.
- (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 B.05*; Gaussian, Inc.: Wallingford, CT, 2004.
- (34) Cossi, M.; Rega, N.; Scalmani, G.; Barone, V. *J. Comput. Chem.* **2003**, *24*, 669–681.
- (35) Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. B* **1998**, *102*, 3257–3271.
- (36) Jang, Y. H.; Goodard, W. A., III; Noyes, K. T.; Sowers, L. C.; Hwang, S.; Chung, D. S. *J. Phys. Chem. B* **2003**, *107*, 344–357.
- (37) Orozco, M.; Luque, F. J. *J. Am. Chem. Soc.* **1995**, *117*, 1378–1386.
- (38) Weigend, F.; Häser, M. *Theor. Chem. Acc.* **1997**, *97*, 331–340.
- (39) Ahlrichs, R.; Bär, M.; Häser, M.; Horn, H.; Kölmel, C. *Chem. Phys. Lett.* **1989**, *162*, 165–169.
- (40) Ahlrichs, R.; Bär, M.; Häser, M.; Horn, H.; Kölmel, C. *Chem. Phys. Lett.* **1998**, *294*, 143–152.
- (41) Schäfer, A.; Huber, C.; Ahlrichs, R. *J. Chem. Phys.* **1994**, *100*, 5829–5835.
- (42) Shapiro, N. D.; Toste, F. D. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 2779–2782.
- (43) Nechaev, M. S.; Rayón, V. M.; Frenking, G. *J. Phys. Chem. A* **2004**, *108*, 3134–3142.
- (44) Kim, C. K.; Lee, K. A.; Kim, C. K.; Lee, B.; Lee, H. W. *Chem. Phys. Lett.* **2004**, *391*, 321–324.
- (45) Hertwig, R. H.; Koch, W.; Schröder, D.; Schwarz, H.; Hrušák, J.; Schwerdtfeger, P. *J. Phys. Chem.* **1996**, *100*, 12253–12260.
- (46) Ziegler, T.; Rauk, A. *Inorg. Chem.* **1979**, *18*, 1558–1565.
- (47) Dewar, M. J. S. *Bull. Soc. Chim. Fr.* **1951**, C71–C79.
- (48) Chatt, J.; Duncanson, L. A. *J. Chem. Soc.* **1953**, 2939–2947.
- (49) Carpenter, J. E.; Weinhold, F. *J. Mol. Struct. (THEOCHEM)* **1988**, *169*, 41–62.
- (50) Carpenter, J. E.; Weinhold, F. *The Structure of Small Molecules and Ions*; Plenum: New York, 1988.
- (51) Barrows, S. E.; Eberlein, T. H. *J. Chem. Educ.* **2005**, *82*, 1329–1333.
- (52) Douglas, J. E.; Rabinovitch, B. S.; Looney, F. S. *J. Chem. Phys.* **1955**, *23*, 315–323.
- (53) Brown, T. J.; Dickens, M. G.; Widenhofer, R. A. *J. Am. Chem. Soc.* **2009**, *131*, 6350–6351.
- (54) Dias, H. V. R.; Wu, J. *Eur. J. Inorg. Chem.* **2008**, 509–522.
- (55) Besler, B. H.; Merz, K. M., Jr.; Kollman, P. A. *J. Comput. Chem.* **1990**, *11*, 431–439.
- (56) Singh, U. C.; Kollman, P. A. *J. Comput. Chem.* **1984**, *5*, 129–145.

Effective Approximation of Molecular Volume Using Atom-Centered Dielectric Functions in Generalized Born Models

Jianhan Chen*

Department of Biochemistry, Kansas State University, Manhattan, Kansas 66506

Received May 12, 2010

Abstract: The generalized Born (GB) theory is a prime choice for implicit treatment of solvent that provides a favorable balance between efficiency and accuracy for reliable simulation of protein conformational equilibria. In GB, the dielectric boundary is a key physical property that needs to be properly described. While it is widely accepted that the molecular surface (MS) should provide the most physical description, most existing GB models are based on van der Waals (vdW)-like surfaces for computational simplicity and efficiency. A simple and effective approximation to molecular volume is explored here using atom-centered dielectric functions within the context of a generalized Born model with simple switching (GBSW). The new model, termed GBSW/MS2, is as efficient as the original vdW-like-surface-based GBSW model, but is able to reproduce the Born radii calculated from the “exact” Poisson–Boltzmann theory with a correlation of 0.95. More importantly, examination of the potentials of mean force of hydrogen-bonding and charge–charge interactions demonstrates that GBSW/MS2 correctly captures the first desolvation peaks, a key signature of true MS. Physical parameters including atomic input radii and peptide backbone torsion were subsequently optimized on the basis of solvation free energies of model compounds, potentials of mean force of their interactions, and conformational equilibria of a set of helical and β -hairpin model peptides. The resulting GBSW/MS2 protein force field reasonably recapitulates the structures and stabilities of these model peptides. Several remaining limitations and possible future developments are also discussed.

1. Introduction

An accurate description of the solvent environment is critical in molecular modeling of biomolecule structure and dynamics. Traditional explicit inclusion of water molecules arguably provides the most detailed description of the solvent but, at the same time, dramatically increases the system size and thus the associated computational cost. The expensive computational cost further hinders extensive optimization of explicit solvent force fields for accurate description of protein conformational equilibria.^{1,2} Instead, implicit treatment of the solvent environment has recently emerged as a powerful alternative to explicit water in biomolecular modeling.³ Implicit solvent essentially aims to capture the mean influ-

ence of solvent molecules on the solute through direct estimation of the solvation free energy, defined as the reversible work required to transfer the solute from vacuum to solution in a fixed configuration. Elimination of the solvent molecules with the implicit treatment substantially reduces the number of atoms needed to be simulated. More importantly, this can now be achieved with only a moderate increase in the computational overhead required for estimating the solvation free energy on-the-fly, such as using continuum electrostatics-based methods including Poisson–Boltzmann (PB) and generalized Born (GB) theories.^{4–6} Such a dramatic reduction in the system size does not come without a loss of detail and certain intrinsic limitations. In general, implicit solvent models may yield considerable disagreement with explicit water simulations in short-range effects when the detailed interplay of a few water molecules

* Phone: (785) 532-2518; fax: (785) 532-7278; e-mail: jianhanc@ksu.edu.

(which are distinct from the bulk water) is important.^{7,8} Such limitations have motivated several attempts to better describe the short-range effects either through empirical corrections⁹ or using hybrid explicit/implicit representations.^{10,11} Implicit solvent might be further limited by the specific methodology for estimating the solvation free energy, as well as the (physical) parameters of the solvation model and underlying protein force field. Nonetheless, the substantial reduction in the computational cost and extension of accessible simulation time scales with implicit treatment of solvent is an important advantage. It not only allows careful optimization of the force field through extensive peptide simulations,^{12–15} but has also opened a door to address many biological problems that are otherwise difficult with explicit solvent.^{4,6}

Among various approaches for implicit treatment of the solvent, GB has been particularly successful for molecular dynamics simulation of biomolecules and especially proteins.⁶ In GB, the total solvation free energy is decomposed into nonpolar and electrostatic contributions:³

$$\Delta G_{\text{solv}} = \Delta G_{\text{elec}} + \Delta G_{\text{np}} \quad (1)$$

Such a decomposition is path-dependent, but it allows both contributions to be related to appropriate (continuum) models of water and is generally more accurate than fully empirical approaches that directly estimate the total solvation free energy from certain solute geometric properties.^{16,17} Given the well-established continuum electrostatics description of water, where the solute is represented as a low dielectric cavity embedded in a featureless, high dielectric solvent medium, the electrostatic solvation free energy can be rigorously calculated by solving the PB equation using finite-difference methods.^{18–20} Alternatively, the GB pairwise approximation can be used to calculate the same quantity.^{21,22}

$$\Delta G_{\text{elec}} = -\frac{1}{2}\tau \sum_{ij} \frac{q_i q_j}{\sqrt{r_{ij}^2 + R_i^{\text{GB}} R_j^{\text{GB}} \exp(-r_{ij}^2 / FR_i^{\text{GB}} R_j^{\text{GB}})}} \quad (2)$$

where r_{ij} is the distance between atoms i and j and q_i is the atomic charge and R_i^{GB} the *effective Born radius* of atom i . $\tau = 1 - 1/\epsilon_s$, with ϵ_s being the solvent (high) dielectric constant. F is an empirical factor whose value may range from 2 to 10, with 4 being the most common one. The GB approximation is much more efficient than PB computationally. At the same time, it provides analytical forces and is thus particularly suitable for molecular dynamic simulations. Accurate calculation of the nonpolar solvation free energy is much more challenging for biomolecules with complex shapes. At present, the nonpolar solvation free energy is either largely ignored or estimated directly from the solvent-accessible surface area (SA) with a phenomenological surface tension coefficient, $\Delta G_{\text{np}} = \gamma S$. With substantial improvement in the electrostatic solvation models, limitations of simple SA models are becoming increasingly important for accurate simulation of protein conformational equilibria.^{23–25}

The effective Born radius, R_i^{GB} , is a key quantity in the GB formalism. It corresponds to the distance between a particular atom and its hypothetical spherical dielectric

boundary, chosen such that the atomic electrostatic self-solvation energy satisfies the Born equation²⁶

$$\Delta G_{\text{elec},i} = -\frac{1}{2}\tau \frac{q_i^2}{R_i^{\text{GB}}} \quad (3)$$

In principle, the “exact” effective Born radii can be calculated from eq 3 using the electrostatic self-solvation energy obtained through the PB theory. It has been shown that, given accurate effective Born radii, the GB approximation of eq 2 closely reproduces the PB electrostatic solvation energy.^{27,28} As such, most of the extensive literature on extensions of the GB theory has been focused on efficient and accurate evaluation of the effective Born radii, and PB-derived $\Delta G_{\text{elec},i}$ or R_i^{GB} have served as standard benchmarks for assessing the numerical accuracy of various GB approximations.^{9,29–41} At present, the GB formalism has reached a mature stage, and many of the latest models are capable of achieving a level of numerical accuracy that approaches that of high-resolution PB calculations.²⁸ Such numerical accuracy provides a necessary basis for recent efforts to optimize various GB-based implicit solvent protein force fields for an accurate description of peptide and protein conformational equilibria.^{12–15}

It should be emphasized that most GB models achieve high numerical accuracy through careful optimization of what can be referred to as “numerical parameters”: parameters specific to a particular GB formalism that are adjusted to maximally reproduce the exact results from equivalent high-resolution PB calculations. Key quantities that need to be reproduced in reference to PB include solvation free energies of small model compounds as well as proteins with various sizes and structures and, most importantly, the effective Born radii (see eq 3) in complex protein environments. These numerical parameters should be distinguished from “physical parameters”, such as the definition of the dielectric boundary (if adjustable), the intrinsic atomic radii for defining the location of the dielectric boundary, and parameters associated with the nonpolar solvation component (e.g., effective surface tension coefficients). These parameters all have well-defined physical meanings and should be optimized to reproduce certain (experimental) physical properties. In particular, optimization of the physical parameters is meaningful when and only when satisfactory numerical accuracy has been achieved. Otherwise, improper cancellation of errors might occur, and this limits the transferability of the optimized implicit solvent force field. It should be noted that there has been some confusion in the literature concerning the optimization of GB implicit solvent, where numerical accuracy is often confused with physical accuracy or physical accuracy is discussed without first establishing the numerical accuracy of the method. Importantly, careful optimization of physical parameters is also necessary for reliable application of various PB methods to biomolecular modeling.

We have previously optimized the intrinsic atomic radii and backbone torsion potentials of the GBSW (generalized Born with simple switching) implicit solvent³⁸ together with the underlying CHARMM param22/CMAP protein force field^{42–45} for reliable simulation of peptide conformational

equilibrium.^{12,13} The optimized force field appears to achieve a reasonable balance of competing solvation and intermolecular interactions and successfully folds a set of diverse model peptides and miniproteins. It has also been applied to model the conformational equilibria of stable and unstable states of several small proteins,^{46–48} including two intrinsically disordered proteins.^{49,50} Nonetheless, application of this force field to folding of larger proteins has not been as satisfactory, and the simplistic SA-based treatment of non-polar solvation has been proposed to be a key limitation.²⁵

The focus of the current work is to explore and address another methodological limitation of GBSW in the description of the solvent–solute dielectric boundary. Specifically, GBSW utilizes an atomic-centered function to smoothly switch between high (water) and low (solute) dielectric regions.^{38,51} Such a smooth dielectric function captures a van der Waals (vdW)-like surface and is computationally efficient and numerically stable. Similar atom-centered dielectric functions have also been widely utilized in many GB (and some PB) models.^{6,28} However, it has been recognized that vdW-like surface definitions result in small, solvent-inaccessible (and thus unphysical) high dielectric pockets, which can lead to significant overestimation of the solvation free energy for larger proteins.^{10,52,53} Instead, the Lee–Richards molecular surface (MS),⁵⁴ defined by rolling a (solvent) probe sphere over the surface of the solute molecule, arguably provides the most appropriate dielectric boundary. Adopting MS in implicit solvent models is very challenging, though, due to a lack of analytical definition, discontinuities with respect to infinitesimal atomic displacements, and sensitivity to grid discretization.³⁷ Several attempts have been made in approximating the original Lee–Richards MS using analytical functions in the context of GB during recent years with various levels of success.^{9,55–57} The GBMV2 (generalized Born using molecular volume) model developed by Lee et al.⁵⁵ has been one of the most successful models in the ability to reproduce PB-derived effective Born radii and total solvation free energies of proteins. Nonetheless, GBMV2 is substantially more expensive than comparable vdW-like surface-based GB models such as GBSW in computational cost.²⁸ More importantly, the sharp molecular surface definition leads to unstable atomic forces and poor energy conservation properties.⁵⁸ A possible remedy is to adopt a smoother dielectric transition, which reduces the ability to approximate the true MS, and at the same time decreases molecular dynamics (MD) time steps to 1–1.5 fs from 2 fs, which further increases the computational cost.⁵⁸ The numerical instability of GBMV2 is an important limitation and has contributed to the difficulty in optimization of the GBMV2 protein force field using similar strategies that prove effective for optimizing GBSW¹³ (Chen, unpublished data).

In the rest of this paper, we will first explore the feasibility and effectiveness of approximating the molecular volume purely on the basis of atom-centered dielectric functions within the framework of GBSW. It will be demonstrated that the new model, termed GBSW/MS2, while as efficient and as stable as the original GBSW model, is able to reproduce the effective Born radii calculated from the MS PB theory

with a correlation of 0.95. More importantly, potentials of mean force (PMFs) of pairwise polar interactions demonstrate that GBSW/MS2 can correctly capture the first desolvation peak, which is a key signature of a true MS-like surface. We then describe further efforts to extensively optimize the GBSW/MS2 protein force field on the basis of solvation free energies, pairwise interactions of a set of amino acid backbone and side chain analogues, and conformational equilibria of several model peptides. At the end, several important remaining limitations and possible directions for further improvement will be discussed.

2. Methods

2.1. Higher Order Corrections to the Coulomb Field Approximation. The original GBSW model³⁸ estimates the atomic electrostatic self-solvation free energy on the basis of the Coulomb field approximation (CFA) with a higher order correction term:³⁷

$$\Delta G_{\text{elec},i} = a_0 \Delta G_{\text{elec},i}^0 + a_1 \Delta G_{\text{elec},i}^1 \quad (4)$$

$$\Delta G_{\text{elec},i}^0 = -\frac{1}{2} \tau q_i^2 \left(\frac{1}{\eta} - \frac{1}{4\pi} \int_{r>\eta_i} \frac{V(\mathbf{r}; \{\mathbf{r}_\alpha\})}{|\mathbf{r} - \mathbf{r}_i|^4} d\mathbf{r} \right) \quad (5)$$

$$\Delta G_{\text{elec},i}^1 = -\frac{1}{2} \tau q_i^2 \left(\frac{1}{4\eta^4} - \frac{1}{4\pi} \int_{r>\eta_i} \frac{V(\mathbf{r}; \{\mathbf{r}_\alpha\})}{|\mathbf{r} - \mathbf{r}_i|^7} d\mathbf{r} \right)^{1/4} \quad (6)$$

where η is an arbitrarily chosen integration starting point less than or equal to the vdW radius of atom i necessary to avoid the singularity at $r = |\mathbf{r} - \mathbf{r}_i| = 0$. The solute interior volume function, $V(\mathbf{r}; \{\mathbf{r}_\alpha\})$, is a function of all atomic positions, $\{\mathbf{r}_\alpha\}$. It is defined by overlapping atom-centered dielectric functions with smooth switching at the solute–solvent boundary:⁵¹

$$V(\mathbf{r}; \{\mathbf{r}_\alpha\}) = 1 - \prod_i H_i(|\mathbf{r} - \mathbf{r}_i|) \quad (7)$$

where the atomic volume exclusion function, $H_i(r)$, is given as

$$H_i(r) = \begin{cases} 0, & r \leq R_i - w \\ \frac{1}{2} + \frac{3}{4w}(r - R_i) - \frac{1}{4w^3}(r - R_i)^3, & R_i - w < r < R_i + w \\ 1, & r \geq R_i + w \end{cases} \quad (8)$$

where R_i is the *atomic input radius* to define the solvent–solute dielectric boundary and $2w$ is the smooth switching length. Both $H_i(r)$ and its first derivative are continuous. The value of $V(\mathbf{r}; \{\mathbf{r}_\alpha\})$ as defined is 1 in the solute interior and gradually decreases to 0 in the solute exterior. The volume integrals in eqs 5 and 6 are evaluated using an efficient numerical quadrature technique.³⁷

The coefficients a_0 and a_1 in eq 4 are two key numerical parameters in GBSW that are parametrized to reproduce the exact $\Delta G_{\text{elec},i}$ computed from PB with the same dielectric boundary definition.⁵¹ The higher order CFA correction term in eq 6 is an empirical one, and the values of $\Delta G_{\text{elec},i}^0$ and $\Delta G_{\text{elec},i}^1$ (or other higher order terms^{37,59}) are highly cor-

related. Such empirical corrections to CFA are important for accurately reproducing the corresponding PB results in both GBSW and GBMV. One way to rationalize the effectiveness of empirical expressions such as eq 4 is that one should be able to rewrite the exact atomic self-electrostatic solvation free energy as an expansion that includes the CFA approximation and a series of higher order (correction) terms, even though in practice only one such correction term appears to be sufficient.^{37,38,55} In analogy, one might suspect that similar sums of a series of correction terms might also provide a means to approximate MS-like surfaces using atom-centered dielectric functions in a purely empirical fashion.

2.2. Numerical Approximation of the Molecular Surface. In this work, we explore a general expression that is directly parametrized on the basis of MS PB results to approximate an MS-like dielectric boundary:

$$1/R_i^{\text{GB}} = D + C_0 A_4 + C_1 A_7 \quad (9)$$

where $A_4 = -2\Delta G_{\text{elec},i}^0/\tau q_i^2$ and $A_7 = -2\Delta G_{\text{elec},i}^1/\tau q_i^2$. This expression differs slightly from an analogous one used in GBMV2,²⁸ in that $1/R_i^{\text{GB}}$ instead of R_i^{GB} is used. This is mainly to reflect the importance of reproducing small R_i^{GB} , as they correspond to cases with larger contributions to electrostatic solvation energetics. Parametrization based on $1/R_i^{\text{GB}}$ is also more suitable compared to those based on $\Delta G_{\text{elec},i}$ (e.g., eq 4). The reason is that the natural spread of atomic partial charges in proteins superficially increases the correlation between values of $\Delta G_{\text{elec},i}$ from GB and PB (via the factor q_i^2), which impedes direct optimization of the ability to mimic MS. Implementation of eq 9 requires minimal changes to the original GBSW module³⁸ available as part of the CHARMM program.^{60,61} This new numerical approximation is referred to as GBSW/MS2. Feig et al. also previously parametrized GBSW directly to reproduce MS PB-derived $\Delta G_{\text{elec},i}$ for the case of $w = 0.2$. This optimization was meant as a quick test of mimicking MS within the GBSW framework and has not been extensively tested. It will be referred to as GBSW/MS hereafter. The GBSW/MS fit is equivalent to assigning $D = 0$, $C_0 = 1.204$, and $C_1 = 0.187$ in eq 9. We note that additional higher order terms could be included in eq 9. However, they do not appear to further improve the goodness of fit, likely because the values of these terms are highly correlated. The effectiveness of approximating an MS-like surface will be examined on the basis of the ability to reproduce MS PB results including atomic effective Born radii and total electrostatic solvation free energies of a protein test set. A key signature of the true MS surface is the inclusion of re-entrant surfaces,⁶² which are manifested as the first major desolvation peaks in the PMFs of pairwise interactions. The ability to describe the first desolvation peak might serve as an important validation of whether such empirical parametrization can sufficiently mimic an MS-like dielectric boundary. Importantly, the proposed reparametrization does not change the underlying smooth definition of the dielectric boundary, and thus, the new model is as efficient and as stable as the original GBSW.

2.3. Optimization of Physical Parameters. Once the numerical accuracy of GBSW/MS2 was established, key physical parameters of the GBSW/MS2 protein force field such as atomic input radii ($\{R_i\}$), a sufficient tension coefficient (γ), and peptide backbone torsion energetics were optimized on the basis of a set of experimental and theoretical properties. The optimization strategy, briefly summarized below, is analogous to what was utilized previously to optimize the original GBSW protein force field.¹³ In principle, other protein force field parameters, particularly Lennard-Jones parameters and atomic partial charges, need to be co-optimized with the new solvation model to achieve full consistency and maximal transferability. However, such an attempt to reparametrize the underlying protein model appears to be highly ambitious. As a compromise and first step, it should be reasonable to focus primarily on directly adjusting the input radii and backbone torsion energetics. An important caveat of such a limited optimization strategy is that one might incorrectly compensate for certain artifacts of the underlying protein model.

2.3.1. Solvation Free Energies and Potentials of Mean Force. The solvation free energies of amino acid side chain analogues are among the few types of experimental data that can be directly used in protein force field parametrization. However, key to an accurate description of peptide conformational equilibria is the ability to capture the delicate balance between sets of competing interactions, i.e., the solvation preference of side chains and backbones versus the strength of (solvent-mediated) interactions between these moieties in a complex protein environment. These two opposing effects are both large and mostly cancel each other. As such, small relative errors in either term might translate into a substantial shift in the balance. Therefore, a more effective approach is to optimize the solvation model directly on the basis of its ability to capture the balance of solvation and intermolecular interactions. Specifically, important physical parameters such as the atomic input radii and surface tension coefficient were systematically optimized to reproduce the strengths of a total of 44 pairwise and three-body interactions among polar and nonpolar model compounds in the TIP3P explicit solvent. A list of all 19 polar PMFs and 25 nonpolar PMFs is provided in the Supporting Information (Figures S1 and S2). Calculation of these TIP3P PMFs has been described in detail in our previous works.^{13,24,25} PMFs in implicit solvent were computed by directly translating the molecules along the axis of interaction. Experimental and TIP3P solvation free energies of amino acid side chain analogues⁶³ were only used for postoptimization validation in this work.

2.3.2. Conformational Equilibria of Model Peptides. After initial optimization of input radii based on PMFs, an iterative procedure was used to empirically fine-tune the solvation parameters together with the peptide backbone torsion energetics, guided by simulation of conformational equilibria of a set of model peptides. The main objective is to balance the solvation model with the underlying CHARMM param22/CMAP protein force field,⁴²⁻⁴⁵ such that both the experimental structures and stabilities of these model peptides can be recapitulated. The model peptides include (1) (AAQAA)₃

(~50% helical at 270 K⁶⁴), (2) GB1p (GEWY DDATK TFTVT E, β -hairpin, 42% folded at 278 K⁶⁵), (3) GB1m1 (GEWY DDATK TATVT E, β -hairpin, 6% folded at 298 K⁶⁶), (4) HP5A (KKYTW NPATG KATVQ E, β -hairpin, 21% folded at 298 K⁶⁶), and (5) GB1m3 (KKWY NPATG KFTVQ E, β -hairpin, 86% folded at 298 K⁶⁶). Consistent with the experimental conditions, the termini of the (AAQAA)₃ peptide were blocked with Ace and NH₂, and all the other peptides were simulated with unblocked termini. The hairpins GB1m1, HP5A, and GB1m3 are derived from the native sequence of the C-terminal β -hairpin (residues 41–56) of the B1 domain of protein G (GB1p), but display reduced or enhanced stability: (unfolded) GB1m1 < HP5A < GB1p < GB1m3 (most folded).⁶⁶ Therefore, these peptide sequences provide a particularly useful control for protein force field optimization.

Heavy reliance on simulation of peptide conformational equilibria is an important aspect of the current optimization strategy. The ability to recapitulate the experimental structures and stabilities of the above model peptides provides key feedback for the parametrization of both the atomic input radii and peptide backbone torsion energetics. An important limitation, however, is the slow convergence of conformational equilibria even for small β -hairpins. In this work, we mainly rely on replica exchange molecular dynamics (REMD), as implemented in the MMTSB Toolset⁶⁷ (available from <http://mmtsb.org>), to improve convergence. While REMD has been shown to be able to provide enhanced conformational sampling for cases with positive enthalpies of activation, important questions remain in the true efficiency and optimal setups for sampling protein conformations in current implicit or explicit solvent protein force fields.^{68–72} In this work, we chose to adopt REMD setups similar to those used in a previous work,¹³ i.e., 16 replicas distributed exponentially within temperatures of 270–500 K. Convergence of simulated conformational ensembles is tested by comparing two independent “control” and “folding” simulations, initially from folded and fully extended conformations, respectively. Such independent simulations from completely different initial coordinates are much more reliable in establishing convergence than simply following the time evolution of the simulations. Additional simulations were also carried out with temperature spans of 270–400 or 270–800 K to seek potential improvement in convergence. Exchanges of simulation temperatures of replicas were attempted every 2 ps, and more frequent exchanges did not appear to improve sampling in any of our tests. The length of the REMD simulation is 20 ns for (AAQAA)₃ and ranges from 50 to 150 ns for GB1p series hairpins. The actual exchange acceptance ratios ranged from about 30% to over 70%. As will be discussed in the Results and Discussion, critical limitations appear to exist in obtaining converged conformational equilibria for several hairpins with GBSW/MS2. However, it is not obvious whether this is reflecting a much greater sampling requirement due to the presence of significant desolvation barriers with the underlying MS-like surfaces or one might be able to fine-tune REMD for individual peptides to improve the sampling efficiency.

Given the significant challenges and expensive computational costs of obtaining converged conformational equilibria of even small model peptides, the iterative optimization relies heavily on manual adjustment of the input radii of important backbone atoms (including amide nitrogen and carbonyl oxygen) and peptide backbone dihedral energetics. Modification of the backbone dihedral energetics was realized using the CMAP dihedral cross-term facility in CHARMM.^{43–45} As a proper balance of secondary structure preference is one of the primary goals, the modifications were focused on the extended (β) and helical regions of the ϕ/ψ space. Stabilization (or destabilization) of particular conformations was achieved by adding cosine-shaped “valleys” (or “humps”) centered at the appropriate ϕ/ψ coordinates to a quantum mechanical CMAP.⁴³

$$\Delta E(\phi, \psi) = \frac{1}{2}k^\alpha[1 + \cos(d^\alpha\pi/r^\alpha)] + \frac{1}{2}k^\beta[2 + \cos(d_1^\beta\pi/r^\beta) + \cos(d_2^\beta\pi/r^\beta)] \quad (10)$$

with

$$d^\alpha = \min[r^\alpha, \sqrt{(\phi - \phi^\alpha)^2 + (\psi - \psi^\alpha)^2}]$$

and

$$d_l^\beta = \min[r^\beta, \sqrt{(\phi - \phi_l^\beta)^2 + (\psi - \psi_l^\beta)^2}] \quad l = 1, 2$$

where the centers and radii are $(\phi^\alpha, \psi^\alpha) = (-75^\circ, -45^\circ)$ with $r^\alpha = 60^\circ$ (for the α -helical region) and $(\phi_1^\beta, \psi_1^\beta) = (-120^\circ, 125^\circ)$ and $(\phi_2^\beta, \psi_2^\beta) = (-150^\circ, 160^\circ)$ with $r^\beta = 45^\circ$ (for parallel and antiparallel β -strand regions). The functional form of eq 10 is purely empirical and mainly serves to reduce the number of adjustable parameters during iterative optimization. It should be noted that the current optimization focuses on describing conformational equilibria, and the strategy of directly targeting local ϕ/ψ regions might have nontrivial consequences on the kinetics of transitions between various conformational states. Another compromise made here is that the details of the unfolded states, such as a prevalence of poly-L-proline II (PPII) helix-like and α_R helix structures,⁷³ are not explicitly considered.

2.3.3. Control Simulation of Peptides and Proteins. Additional control simulations were also carried out for a few proteins and protein complexes of various sizes and folds to validate the stability of these proteins in the optimized GBSW/MS2 force field. These systems include the B1 domain of protein G (mixed helix/ β , PDB 3gb1), the apo form of dihydrofolate reductase (mixed helix/ β , PDB 1rx2), and the complex between domains of CREB binding protein (CBP) and the activator of thyroid and retinoid receptors (ACTR) (helical, PDB 1kbh). In addition, the conformational equilibrium of a short polyaniline peptide, Ala₅, is calculated using a 40 ns replica exchange simulation to more directly examine the ability of the optimized GBSW/MS2 force field to describe unfolded protein states. Eight replicas spanning 300–500 K were used, and exchanges of replica temperatures were attempted every 2 ps. Consistent with the experimental conditions (carried out at pH 2), both the N- and C-termini of Ala₅ are protonated. NMR scalar coupling

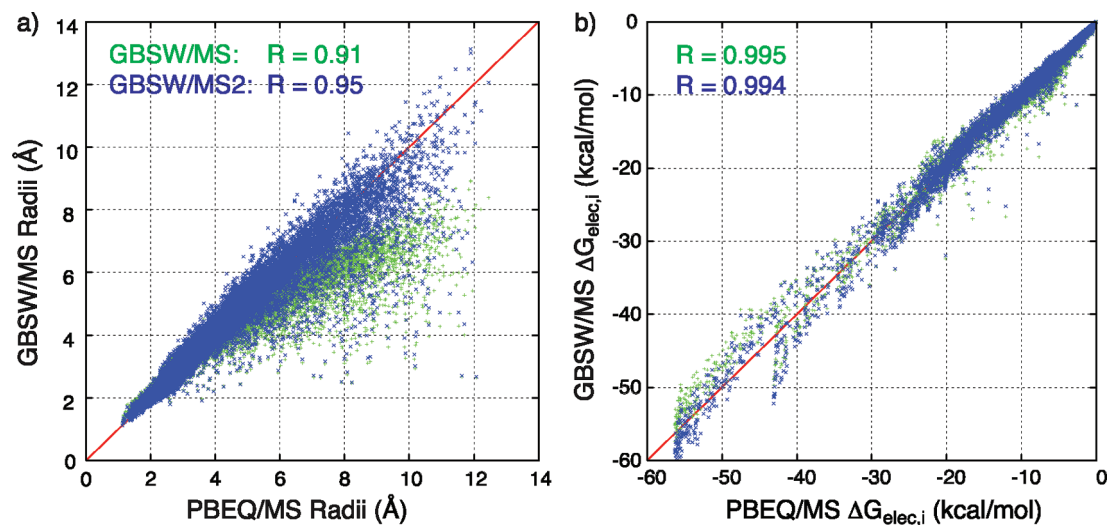


Figure 1. Comparison between PB and GB effective Born radii (a) and atomic electrostatic self-solvation free energies (b) for all atoms of a set of 22 small proteins. Green and blue points correspond to GBSW/MS and GBSW/MS2 results, respectively.

constants are calculated on the basis of the Karplus equation using the same parameters as in the NMR study,⁷³ specifically ${}^3J(\text{H}_\text{N}, \text{H}_\alpha) = 7.09 \cos^2(\phi - 60^\circ) - 1.42 \cos(\phi - 60^\circ) + 1.55$ and ${}^2J(\text{N}, \text{C}_\alpha) = -0.66 \cos^2 \psi_{i-1} - 1.52 \cos \psi_{i-1} + 7.85$.

3. Results and Discussion

3.1. Numerical Parametrization: Effective Born Radii and Total Protein Solvation Free Energies. Numerical parameters in eq 9 were obtained by minimizing the root-mean-square error between GB and PB results for all atoms in a set of 22 small proteins (test set 1 of Feig et al.,²⁸ also see Table 1 of the Supporting Information). The sizes of these proteins range from 37 to 98 residues. The MS PB results were computed using the PBEQ module⁵¹ in CHARMM with a grid spacing of 0.21 Å and a probe radius of 1.4 Å. The switching length in eq 8 is set to $2w = 0.4$ Å. A previously optimized set of input radii^{74,75} (hereinafter referred to as Nina's radii) was used in both GB and PB calculations at this stage. The best fit was achieved with $D = -0.0505$, $C_0 = 1.437$, and $C_1 = 0.1631$, with a correlation coefficient of $R = 0.98$ between GB- and PB-derived $1/R_i^{\text{GB}}$ values. Figure 1 compares the PB and GB effective Born radii and electrostatic self-solvation free energies. The correlations of the self-solvation free energies of GBSW/MS and GBSW/MS2 with PB are similar, and both are somewhat misleadingly high, with $R \approx 0.995$ due to the natural spread of atomic partial charges. This reinforces the notion that self-solvation free energies are not sufficiently sensitive for numerical parametrization of GB. The effective Born radii are the most important quantities to be estimated accurately in a GB model. GBSW/MS2 improves the correlation between GB- and PB-derived effective Born radii to $R = 0.95$ compared to $R = 0.91$ for GBSW/MS. While such a correlation is still substantially lower than that of GBMV2 ($R \approx 0.99$), it is a significant improvement over that of $R = 0.811$ achieved by a GSGB model developed by Yu and co-workers.⁵⁶ The two other previous attempts to

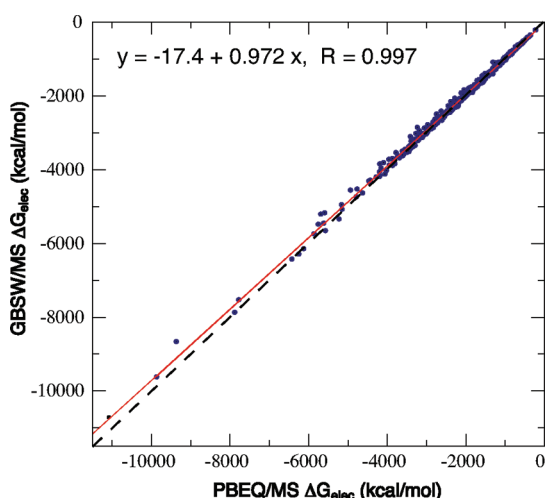
incorporate MS-like surfaces in GB^{9,57} do not report numerical values for the correlation between GB and PB effective Born radii. Nonetheless, it might be estimated that the GBn model described by Morgan and co-workers does not perform better than GSGB in reproducing PB-derived effective Born radii (on the basis of Figure 3 of ref 57).

The effectiveness of numerical parametrization in reproducing the total electrostatic solvation free energies was examined using both test set 1 and a larger set of 611 proteins (test set 2 of Feig et al.²⁸). Test set 2 contains nonhomologous, single-chain proteins that range from small protein fragments to large ones with over 800 residues and cover diverse native folds. The total electrostatic solvation free energies for all 22 proteins in test set 1 (which is the training set for numerical parametrization) are provided in Table 1 of the Supporting Information. Nina's radii were used in these calculations. The average error improves to less than 2% for GBSW/MS2 compared to that of over 6% for GBSW/MS. The maximum relative error of 9.6% was observed for a small 46-residue protein (PDB 1cbn). It is not clear why GBSW/MS and GBSW/MS2 appear to have particular difficulty for this protein. The maximum relative error is about 4% for the rest of the training set. In Figure 2, we compare the total solvation free energies calculated from GBSW/MS2 and PBEQ for test set 2. As will be shown later in this section, the CHARMM param22 vdW radii are nearly optimal for GBMV2. To make direct comparison to GBMV2 and also to test the sensitivity of the above numerical parametrization to (small) changes in input radii, the CHARMM param22 vdW radii were used in the calculations shown in Figure 2. The apparent correlation between PB and GB results is excellent, with a correlation coefficient of $R = 0.997$. The maximum and average absolute and relative errors are 708.2 and 47.7 kcal/mol, compared to those of 482.8 and 28.1 kcal/mol of GBMV2 (with default options and SHIFT -0.102 SLOPE 0.9085 P6 8). Interestingly, the GBSW/MS2 solvation free energies are actually better correlated with the GBMV2 results than with PBEQ/MS,

Table 1. Optimized Atomic Input Radii for GBSW/MS2 and GBMV2 in Comparison with the Original CHARMM param22 vdW and Nina's Radii^a

group	atom ^b	vdW	GBSW/		
			Nina	MS2	GBMV2 ^c
hydrogen		varies	0.0	0.0	–
backbone	C	2.06	2.04	2.06	2.06
	O	1.70	1.52	1.75	1.70
	N	1.85	2.23	1.95	1.85
	CA	2.06	2.86	2.86	–
side chains					
all ^d	CB	2.175	2.67	2.80	–
all ^d	CD, CG	2.275	2.46	2.50	–
Lys	NZ	1.85	2.13	2.13	1.85
Arg	CE	2.175	2.80	2.60	2.80
	NH*	1.85	2.13	2.05	1.80
	NE	1.85	2.13	2.10	1.80
Glu ^e	CZ	2.00	2.80	2.60	2.80
	OE*	1.70	1.42	1.78	1.70
Gln ^f	NE2	1.85	2.15	2.10	1.95
	OE1	1.70	1.42	1.85	1.70
His ^g	ND1, NE2	1.85	2.31	2.10	1.85
	CD2, CE1	1.80	1.85	2.20	–
Trp	NE1	1.85	2.40	2.05	1.85
	C* (ring)	1.8–2.0	1.78	2.30	–
Tyr	OH	1.77	1.85	1.73	1.77
Ser, Thr	OG*	1.77	1.64	1.70	1.77
Pro	CB, CG, CD	2.175	1.98	2.20	–
Phe, Tyr	C* (ring)	1.99	2.00	2.30	–
methyl carbons ^h		2.06	2.44	2.70	–

^a The CHARMM param22 vdW radii are used for all atoms not shown in this table. All values are in angstroms. ^b * is a Wild card character. ^c Entries with “–” denote cases where the radii have not been optimized for GBMV2 and the default CHARMM para22 vdW radii are used. ^d Unless otherwise specified. ^e Also for OD* of Asp and carbonyl oxygen atoms in the charged C-terminus. ^f Also for ND2 and OD1 of Asn. ^g Also for protonated His (Hsp). ^h Except CB of Ala.

**Figure 2.** Comparison of the total electrostatic solvation free energies computed using GBSW/MS2 and PBEQ for all 611 proteins in test set 2. The dashed line plots the diagonal, and the red line corresponds to the best linear fit. The CHARMM para22 vdW radii were used as the input radii in both GB and PB calculations.

with $R = 0.999$. We note that, while the average error of GBMV2 is similar to what was previously reported,²⁸ the maximum GBMV2 error is higher in the current calculations. This might be attributed to slight difference in the PBEQ setup (Feig et al.²⁸ used a slightly larger grid of 0.25 Å). In

addition, several of our PBEQ calculations of the total self-solvation free energies appeared to be unstable, where multiple calculations with different protein orientations failed to yield similar results and some of them even completely failed in reaching convergence within the maximum number of iterations. Changing grid specifications and other PBEQ parameters did not appear to resolve such occasional instabilities in a consistent fashion. Nonetheless, the average error and correlation coefficient reported above are minimally impacted by the presence of several less reliable PB results.

3.2. Initial Optimization of Input Radii: PMFs and Solvation Free Energies. Initial calculations using various sets of input radii suggested that neither Nina's radii^{74,75} nor a modified set optimized for GBSW¹³ offers an obvious advantage as the starting point for optimizing the GBSW/MS2 protein force field. Instead, initial optimization of the GBMV2 protein force field suggested that the CHARMM param22 vdW radii are nearly optimal for GBMV2. Therefore, the CHARMM param22 vdW radii were specified as the default for GBSW/MS2, and input radii for selected atoms were systematically optimized to reproduce the stabilities of pairwise and multibody polar and nonpolar interactions among backbone and side chain analogues in various poses in TIP3P. Even with nearly 50 PMFs, the parametrization appears to be underdetermined, and many radius sets similarly reproduce the TIP3P results in terms of the root-mean-squared deviation (rmsd) of the stabilities. The input radius set with the fewest modifications from the initial values was chosen to avoid overparametrization. The input radii of key backbone atoms were further co-optimized with the peptide backbone torsion energetics on the basis of peptide simulations (see the next section). We have also optimized the input radii for peptide backbone and polar side chains for GBMV2 to establish a fair benchmark for assessing GBSW/MS2. The final optimized input radii are summarized in Table 1. The associated CHARMM input files for setting up the atomic input radii for GBSW/MS2 and GBMV2 are provided in the Supporting Information. With these radii, GBSW/MS2 is able to reproduce the stabilities of polar pairwise interactions in TIP3P to about 1.1 kcal/mol rmsd (excluding two problematic pairs noted below) and those of nonpolar interactions to about 0.42 kcal/mol rmsd. The strengths of all interactions studied in this work in TIP3P, GBSW, GBMV2, and GBSW/MS2 are summarized Figures S1 and S2 of the Supporting Information.

Figure 3 compares the PMFs of nine representative pairwise interactions between various polar moieties using sets of optimized radii specific to each of the three implicit solvent models compared. These PMFs clearly demonstrate the ability of GBSW/MS2 to capture both the location and magnitude of the first desolvation peaks in the TIP3P PMFs to a degree that is comparable to that of GBMV2. This is in contrast to the original GBSW model, where the desolvation peaks are largely absent for all the pairs examined. A correct description of the desolvation peaks strongly supports that the proposed empirical parametrization is indeed capable of capturing a realistic MS-like solute–solvent boundary at the atomic level, despite the use of atomic-centric dielectric functions. Through input radius optimization, GBMV2 and

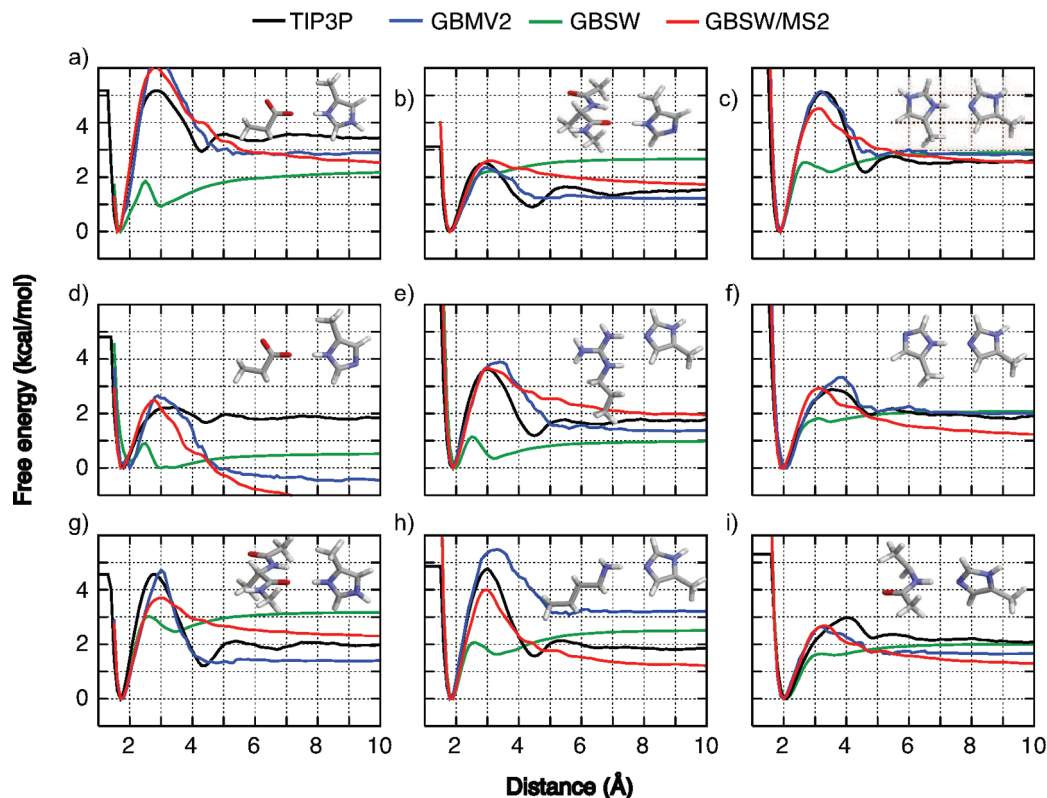


Figure 3. Comparison of the PMFs of nine representative pairwise interactions between polar moieties: (a) Hsp–Glu (hpe), (b) His–backbone carbonyl (hbco), (c) Hsp–His (hhp), (d) His–Glu (he), (e) His–Arg (hr), (f) His–His (hh), (g) Hsp–backbone carbonyl (hpbco), (h) His–Lys (hk), and (i) His–backbone amide (hb). All PMFs are aligned at the contact minimum. Details of the TIP3P PMF calculations were described in a previous paper.¹³ PMFs in GBSW were computed using a previously optimized input radius set,¹³ while those in GBSW/MS2 and GBMV2 were calculated using the optimized radii listed in Table 1.

GBSW/MS2 are able to reproduce the TIP3P stabilities of most pairwise interactions well. However, both GBMV2 and GBSW/MS2 appear to have difficulty in modeling Glu (and Asp), particularly its interaction with His (Figure 3d) or Lys (see Figure S1 in the Supporting Information). There is an unusual sensitivity to the choice of input radii, and no set was able to reproduce all related PMFs. The specific reason is not fully understood at this point. Interestingly, the final optimized radii yield accurate solvation free energies for side chain analogues of Asp/Glu (kcal/mol): $-79.95/-78.03$ (GBMV2) and $-81.02/-78.76$ (GBSW/MS2) compared to $-80.62/-80.54$ (TIP3P⁷⁶) and $-80.65/-79.12$ (experimental⁷⁶). The rms deviation from the TIP3P stabilities for all pairs except ek and he (see Figure S1) is about 1 kcal/mol for both GBSW/MS2 and GBMV2. Closer inspection reveals that GBMV2 PMFs tend to be rugged, which will lead to unstable forces and, consequently, numerical instability and poor energy conservation as previously observed in GBMV2 protein simulations.⁵⁸ The PMFs in GBSW/MS2 also contain small fluctuations. However, these fluctuations are smooth, and the atomic forces from GBSW/MS2 are as stable as in GBSW. Furthermore, as shown in Figure 4, the ability of GBSW/MS2 to reproduce the experimental solvation free energies of amino acid side chains is not compromised even though the input radius optimization was based on PMFs alone.

Previous analyses have strongly suggested that SA-based models have important limitations in describing the protein conformational dependence of nonpolar solvation.^{23,24} In

particular, it has been argued that both the solvent screening of the solute–solute dispersion interactions and length-scale dependence of hydrophobic solvation need to be properly described.²⁵ Ongoing development of more sophisticated nonpolar solvation models is beyond the scope of the current work. Nonetheless, it is important to examine the ability of the simple SA model in reproducing the TIP3P PMFs of nonpolar interactions. The results of all 25 nonpolar interactions examined in this work are summarized in Figure S2 of the Supporting Information. While there remains a systematic underestimation of the strengths of three-body hydrophobic associations (e.g., fff and lll interactions in Figure S2) general to simple SA models,²⁴ GBSW/MS2 with SA can be parametrized to reproduce the TIP3P results to an overall rmsd of 0.42 kcal/mol. Figure 5 plots four representative nonpolar PMFs calculated in TIP3P, GBSW, GBMV, and GBSW/MS2 solvents. All three implicit solvents are able to more or less capture the overall stabilities of these interactions. However, regardless of adopting vdW or MS-like surfaces, none of these (SA) models are able to reproduce the desolvation peaks observed in TIP3P. This further illustrates the insufficiency of SA-based models to capture the delicate balance between cavitation (creating a hydrophobic cavity within the solvent) and solute–solvent dispersion interactions and, particularly, the conformational dependence of this balance.

3.3. Conformational Equilibria of Model Peptides.

Examination of interactions between backbone and side chain moieties described above clearly demonstrates the challenge in

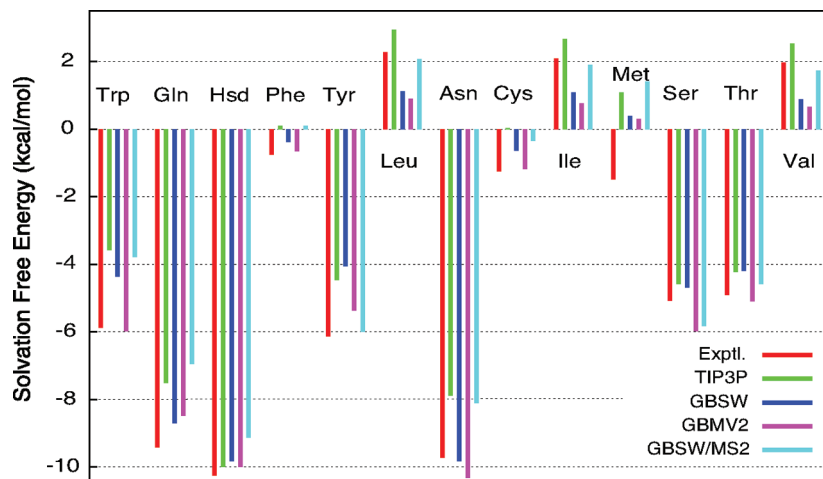


Figure 4. Experimental and calculated total solvation energies of amino acid side chain analogues. The experimental and TIP3P values were taken from ref 63. The GBSW results were computed using a previously optimized set,¹³ and the GBMV2 and GBSW/MS2 results were computed using the radii listed in Table 1. All model compounds have the default geometries as defined in the CHARMM param22 force field.⁴² The rms deviations from the experimental results are 1.41, 1.11, 1.07, and 1.34 kcal/mol for TIP3P, GBSW, GBMV2, and GBSW/MS2, respectively.

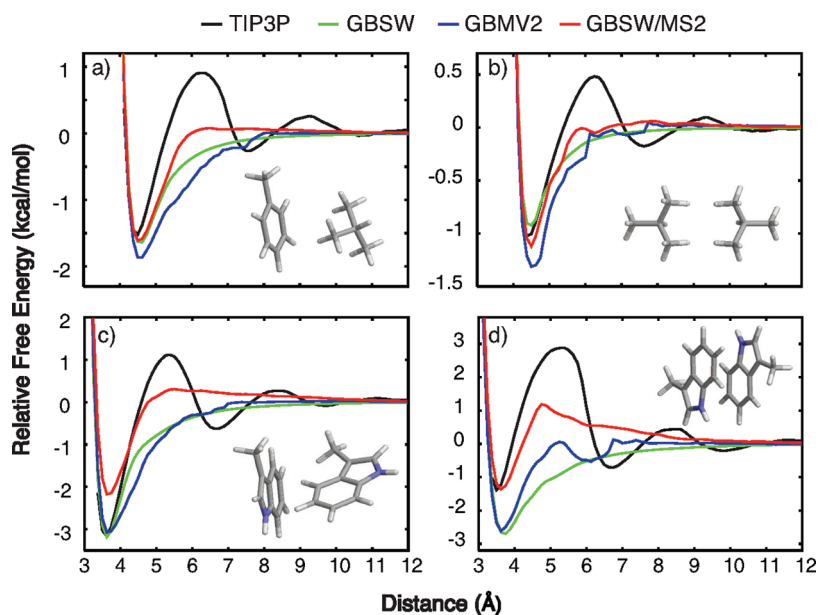


Figure 5. Comparison of the PMFs of four representative pairwise interactions between nonpolar moieties: (a) Phe–Leu (fl), (b) Leu–Leu (head-to-head, ll_h), (c) Trp–Trp (edge-to-face, ww_etf), and (d) Trp–Trp (antiparallel displaced, ww_apd). All PMFs are aligned at the largest separation distances. Details of the free energy protocol for calculating these TIP3P PMFs were described in previous papers.^{24,25} PMFs in GBSW were computed using a previously optimized input radius set,¹³ while those in GBSW/MS2 and GBMV2 were calculated using the radii listed in Table 1.

accurately balancing competing interactions even at the small-molecule level. It is thus important to achieve sufficient cancellation of errors at peptide and protein levels, such as via iterative optimization of important protein parameters, particularly backbone torsional energetics, together with the solvation model. A large number of folding and control REMD simulations (>100) of the model peptides were carried out to explore various parameter sets. The final optimized GBSW/MS2 protein force field achieves reasonable success in balancing the secondary structural propensities and in recapitulating the experimental structure and stabilities of the five model peptides used in this work, but nonetheless with important limitations (detailed later in the discussion). Figure 6 examines the average helicity of

(AAQAA)₃ as a function of temperature for several representative combinations of backbone input radii and CMAP adjustments. The calculated helicity appears to converge well with 20 ns REMD simulations, as illustrated by comparing results from two independent simulations for one of the parameter sets explored (thick and thin solid red traces in Figure 6). Clearly, approximation of an MS-like dielectric boundary significantly increases the intrinsic propensity to form an ideal α -helix. For example, GBSW/MS2 yields an extremely stable helix with a melting temperature of ~ 500 K (solid green trace in Figure 6), even though the corresponding backbone hydrogen-bonding strength estimated from a modified alanine dipeptide dimer model is only 1.55 kcal/mol, weaker than that in either TIP3P

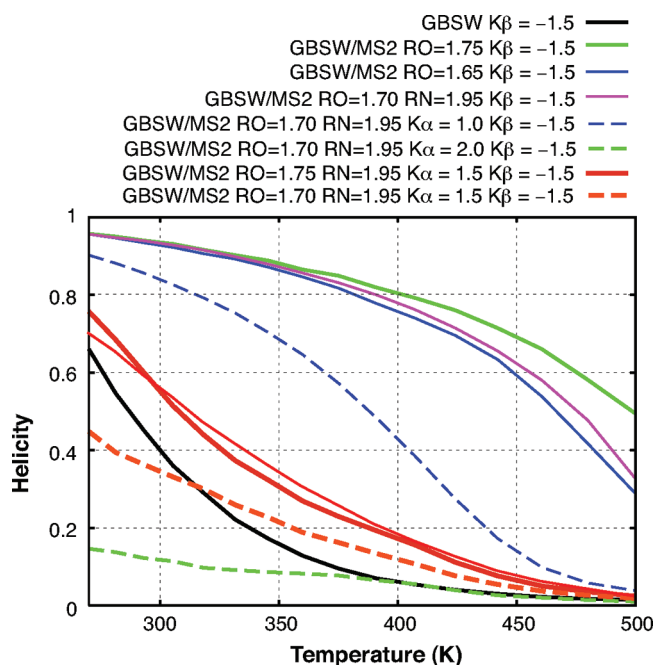


Figure 6. Total helicity of (AAQAA)₃ as a function of temperature in GBSW/MS2 implicit solvent with different combinations of peptide backbone torsion modifications and input radii. The result calculated using a previously optimized GBSW protein force field¹³ is also shown for comparison. The helicity was calculated from the averaged frequency of 1–4 hydrogen bonding, defined by $d_{O_p,HN_{i+4}} \leq 2.6$ Å, during the second halves of 20 ns REMD simulations included in computing the averages. The backbone torsion modifications were implemented using eq 10 with various k^α and k^β values. RO and RN denote the input radii for backbone carbonyl oxygen atoms and amide nitrogen, respectively. RN = 2.05 Å when not explicitly specified. The stabilities of the backbone hydrogen bonding for a modified alanine dipeptide dimer¹² are (kcal/mol) 1.9 ± 0.3 (TIP3P), 1.80 (GBSW), 1.55 (GBSW/MS2, RO = 1.75 Å), 0.94 (GBSW/MS2, RO = 1.65 Å), 1.44 (GBSW/MS2, RO = 1.75 Å, RN = 1.95 Å), and 1.1 (GBSW/MS2, RO = 1.70 Å, RN = 1.95 Å). Results from two independent REMD simulations are shown for GBSW/MS2 with the final selected parameters to illustrate convergence (thick and thin solid red traces).

(1.9 ± 0.3 kcal/mol) or the previously optimized GBSW solvent (1.8 kcal/mol).¹³ Similar observations can be made with GBMV2, as illustrated in Figure S3 of the Supporting Information. This dramatic difference between two types of dielectric boundaries might be rationalized by a significant reduction of internal high dielectric regions (and thus reduced dielectric screening) with an MS compared to a vdW-like surface. Importantly, the intrinsic helical propensity with the MS is strong and cannot be overcome by tuning the backbone hydrogen-bonding strength alone. For example, the stability of (AAQAA)₃ remains grossly overestimated even if one reduces the backbone hydrogen-bonding strength (in the model dimer) to less than 1 kcal/mol (solid blue trace in Figure 6). Instead, it is necessary to modify the peptide backbone torsion energetics, such as by directly reducing the stability of helical conformations. It turns out that k^α up to 2 kcal/mol is sufficient to completely destabilize the helical conformation. The optimal choice appears to be $k^\alpha = 1.5$ kcal/mol and $k^\beta = -1.5$ kcal/

mol (this choice of k^β was based on further simulations of β -hairpins). Given this backbone torsion modification, input radii of key backbone atom types (carbonyl oxygens and amide nitrogens) can be slightly adjusted to fine-tune the balance between helical and random coil/extended conformations, as illustrated by the solid and dashed red traces in Figure 6.

Disappointedly, further fine-tuning of the backbone input radii and ϕ/ψ torsion energetics based on conformational equilibria of the GB1p series hairpins has been met with much greater difficulty compared to our previous efforts in optimizing the GBSW protein force field.¹³ The primary challenge is an apparent inability to achieve convergence on conformational equilibria of GB1p-derived hairpins, even with REMD simulations up to 150 ns. Figure 7 compares the energetic and structural properties of the lowest temperature ensembles (at 270 K) from 50 ns control and folding simulations of GB1m3, the most stable hairpin in the series. While several replicas did successfully fold during the folding simulation (panel c), control and folding runs do not sample similar regions of the conformational space, as reflected in the probability distributions of the number of native hydrogen bonds (panel d). There appear to be significant energetic barriers separating the fully folded basin from nearly folded and unfolded regions, such that the low-energy basin predominantly sampled in the control simulation is not visited during the folding run (panels a and b). One can speculate that, to arrive at (or escape) the fully folded native basin, multiple (polar) interactions need to be rearranged at the same time, which likely involves significant collective (desolvation) barriers due to the MS-like underlying dielectric boundary. In addition, conformational diffusion is expected to be slower in GBSW/MS2 (or GBMV2) than in GBSW due to the presence of a desolvation barrier, which also makes sampling more difficult in general.

Similar difficulty was experienced in achieving convergence in conformational equilibria for other GB1p hairpins. The control REMD simulations systematically overestimate the stabilities, while the folding runs systematically underestimate the stabilities. This is clearly reflected by comparing the probability distributions of the number of native hydrogen bonds in Figure 8. HP5A is the single case where a certain level of apparent convergence was achieved. Similar to previous simulations with GBSW,¹³ GB1p is the most challenging sequence to simulate with a more flexible turn (DATK vs PATG in HP5A and GB1m3). No replica successfully folded in any of the multiple 50 ns REMD folding simulations (e.g., see “folding-1” in Figure 8a). The only folding event was observed during a 150 ns REMD simulation with a smaller temperature range of 270–400 K (“folding-2” in Figure 8; the folding event occurred around the 50 ns mark during the simulation). For GB1m1, no replica ever folded during multiple 50 ns REMD folding simulations with various temperature ranges. However, several replicas remained folded even when the simulation was extended to 100 ns (with a temperature range of 270–500 K) (see Figure 8d). Again, the difficulty in achieving convergence is clearly related to a significantly more rugged energy landscape with an MS-like underlying dielectric boundary. It also reflects important limitations of standard REMD in enhancing conformational sampling and re-emphasizes the importance of careful

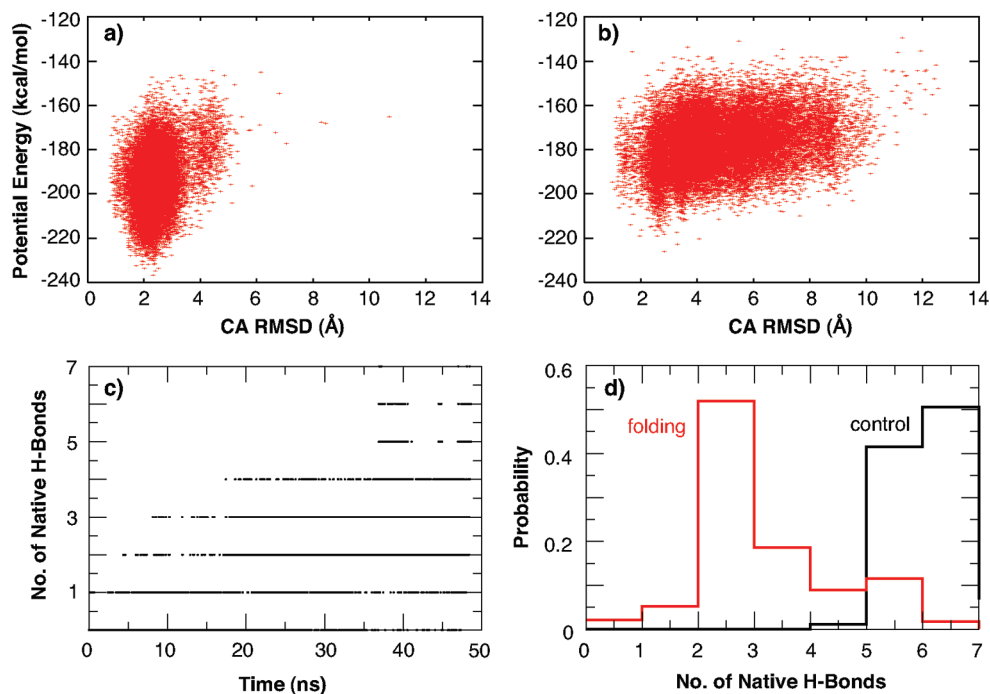


Figure 7. REMD simulations of the gb1m3 hairpin. Potential energy and C_{α} rmsd scatter plots for all conformations sampled at 270 K from (a) control and (b) folding REMD simulations, (c) the number of native hydrogen bonds as a function of simulation time, and (d) probability distributions of the number of native hydrogen bonds at 270 K. The probability distributions were computed from the last 20% (10 ns) of 50 ns REMD control and folding simulations. GBSW/MS2 was used with the atomic input radii as specified in Table 1, together with a modified CMAP (eq 10 with $k^{\alpha} = 1.5$ kcal/mol and $k^{\beta} = -1.5$ kcal/mol). The native hydrogen bonds include (in protein GB1 residue numbering) E42(N)–T55(O), E42(O)–T55(N), T44(N)–T53(O), T44(O)–T53(N), D46(N)–T51(O), D46(O)–T51(N), and D47(O)–K50(N).

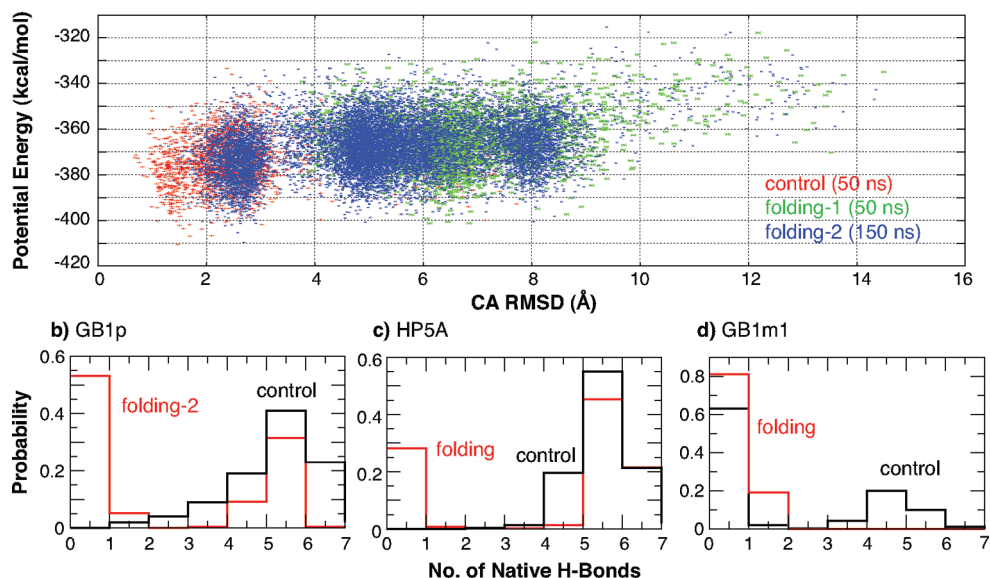


Figure 8. REMD simulations of GB1p, HP5A, and GB1m1: (a) potential energy and C_{α} rmsd scatter plots for all conformations sampled at 270 K from multiple control and folding simulations of GB1p; (b–d) probability distributions of the number of native hydrogen bonds at 270 K. Details of the simulation and analysis are provided in the caption of Figure 7, except as otherwise noted here. Two folding REMD simulations are shown for GB1p, where the temperature range is 270–500 K for “folding-1” and 270–400 K for “folding-2”. The folding simulation of HP5A was also carried out with a temperature range of 270–400 K. The length of the GB1m1 control simulation was 100 ns.

examination of convergence by independent runs initiated from distal points in the conformational space. Nevertheless, results from these folding and control simulations suggest that the optimized GBSW/MS2 force field provides a reasonable balance between helical and coil/extended conformations, being able

to fold both helices and β -hairpins. More importantly, albeit inconclusive due to a lack of satisfactory convergence, results shown in Figures 7 and 8 indicate that GBSW/MS2 roughly captures the trend in stabilities of GB1p series hairpins, $GB1m3 > GB1p \approx HP5A > GB1m1$.

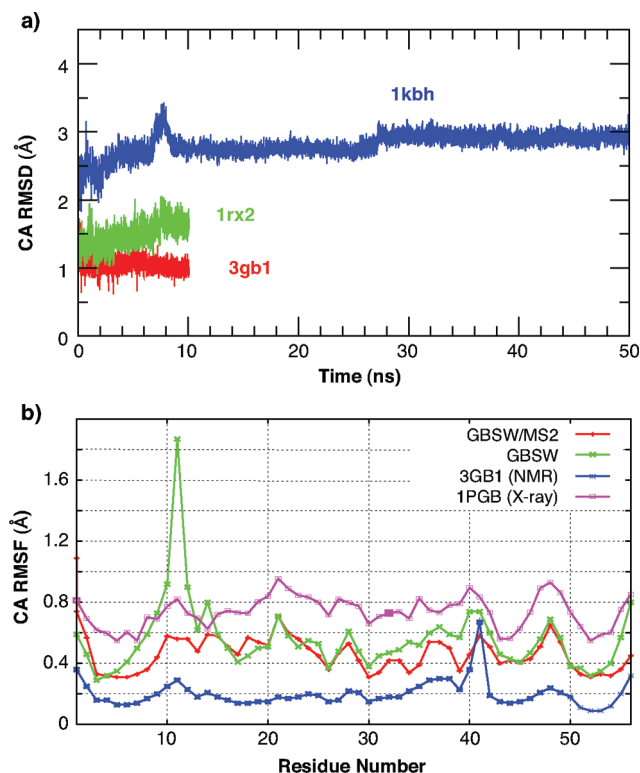


Figure 9. (a) C_{α} rmsd from the PDB structures during 10–50 ns simulations of two proteins (1rx2 and 3gb1) and one protein complex (1kbh) at 300 K and (b) C_{α} RMSF calculated from the second halves of 10 ns MD simulations of 3gb1 in GBSW/MS2 and GBSW implicit solvents in comparison with values calculated from the NMR and X-ray PDB structures. The disorder termini in complex 1kbh were not included in the rmsd calculation. The C_{α} RMSF of PDB 3gb1 was calculated using all 32 models of the NMR ensemble, and that of the X-ray structure (PDB 1pgb) was converted from the B factors using the relationship $B \text{ factor} = 8\pi^2(\text{RMSF})^2/3$.

3.4. Control Simulation of Peptides and Proteins.

Control simulations of two proteins and a protein complex were also carried out to confirm the suitability of simulating the native states using the optimized GBSW/MS2 force field, particularly considering that the previous parametrization (GBSW/MS) appears to have a problem stabilizing several proteins. As shown in Figure 9, all these systems remain stable throughout the simulation time scales up to 50 ns for the CBP/ACTR complex (PDB 1kbh). Note that both CBP and ACTR domains are intrinsically disordered in unbound states and the complex is believed to retain substantial flexibility.^{77,78} In Figure 9b, we compare the root-mean-square fluctuations (RMSFs) computed from 10 ns simulations in GBSW/MS2 and GBSW together with those obtained from NMR and X-ray PDB structures. Even though the first hairpin turn (near residue 11) appears to be substantially more flexible in GBSW than in GBSW/MS2, there is little experimental evidence for this from either the B factors or the NMR ensemble. Interestingly, the RMSF profile from the GBSW/MS2 simulation better tracks the one converted from the B factors, with a correlation of 0.66 compared to that of 0.41 for GBSW. These control simulations suggest that GBSW/MS2 provides a suitable alternative to GBMV2

(or explicit solvent) for protein simulations with enhanced stability and computational efficiency.

The ability of the GBSW/MS2 force field to describe unfolded protein states is examined by comparing the calculated NMR scalar coupling constants with experimental values for Ala₅. The results are summarized in the Supporting Information, Figure S4. Deviations from the experimental values are comparable to previous analysis of other force fields.^{73,79,80} Using the same criteria for classifying the backbone (ϕ , ψ) space as those used by Best et al.,⁸⁰ the populations of α , β , and PPII regions are 7.7%, 69%, and 22% at 300 K, respectively. Interestingly, the calculated helical population using the optimized GBSW/MS2 force field is one of the smallest among all force fields reported (e.g., compared to 41.5% from CHARMM27/CMAP with TIP3P water) and is right within the range of helicities estimated by reweighting various basins on the basis of NMR scalar coupling constants.⁸⁰ However, the β population appears to be significantly overestimated, at the expense of lower PPII conformation. This is likely a consequence of the optimization strategy that directly targets the balance between α/β secondary structure folding propensities. Additional attention to the details of unfolded states will need to be addressed in future studies.

4. Conclusions

MS is widely considered as the most proper choice for representing the solute–solvent boundary in continuum electrostatics-based implicit solvent models such as GB and PB, yet most GB models are based on atom-centered dielectric functions that describe vdW-like surfaces for computational efficiency and numerical stability. A surprisingly simple, yet effective, approximation to true MS is described here in the context of GBSW implicit solvent. By directly optimizing appropriate *numerical parameters*, the sum of a series of CFA and higher order terms was shown to be capable of nicely approximating the true MS. The new model, termed GBSW/MS2, is able to reproduce the exact MS PB-derived effective Born radii with a correlation of 0.95, apparently second only to the GBMV2 model among the few existing MS-based GB models. More importantly, examination of the PMFs of pairwise polar interactions demonstrates that desolvation barriers, a key signature of a true MS, are properly captured in GBSW/MS2. GBSW/MS2 fully retains the computational efficiency and numerical stability of the original GBSW model, allowing one to extensively optimize the *physical parameters* to obtain a balanced GBSW/MS2 protein force field. The optimization was guided by PMFs of over 40 polar and nonpolar interactions and relied extensively on direct simulation of the conformational equilibria of a set of carefully chosen helical and β -hairpin peptides. The key physical parameters optimized include the atomic input radii for specifying the location of solute–solvent boundary and peptide backbone torsion energetics. The final optimized GBSW/MS2 protein force field not only satisfactorily captures the delicate balance between solvation and intramolecular interactions on the model compound level compared to the TIP3P solvent, but also appears to reasonably describe the balance between helical and coil/ β conformations on the model level.

Nonetheless, several important limitations have prevented one from further fine-tuning the GBSW/MS2 force field. In particular, the presence of desolvation peaks due to MS significantly increases the ruggedness of the underlying potential energy surface and hinders expedient conformational sampling. Satisfactory convergence in simulated conformational ensembles could not be achieved for the β -hairpins using REMD simulations up to 150 ns with various temperature ranges. In contrast, good convergence was achieved for the same set of hairpins within 30–50 ns for GBSW with a vdW-like underlying dielectric boundary.¹³ This also highlights the importance of improved understanding of the efficacy of REMD⁷¹ and developing other enhanced sampling techniques^{81,82} in development and optimization of (implicit solvent) protein force fields. The current work focuses on the electrostatic component of the solvation model, and the nonpolar solvation free energy is still estimated using the simple SA model. While reasonable success can be achieved in reproducing stabilities of pairwise and multibody nonpolar interactions in TIP3P, it must be emphasized that SA-based models have important limitations in capturing the conformational dependence of solvation.^{23,25} This deficiency is partially reflected in the lack of fine features in PMFs of nonpolar interactions, using either MS or vdW-like dielectric boundaries. Development of more sophisticated nonpolar solvation models will be necessary to achieve a fully balanced implicit-solvent-based force field for reliable simulation of peptide and protein conformational equilibria. This is a formidable task and will also depend critically on further improvement in the underlying protein models.^{1,2,83} The GBSW/MS2 approximation introduced here not only provides a viable alternative for modeling protein structure and interactions, but is also expected to constitute an important step toward the development of an efficient, stable, and fully balanced GB-based implicit solvent protein force field.

Acknowledgment. I am in debt to Dr. Wonpil Im for providing the protein test sets used in this work, and I thank Drs. Wonpil Im and Debabani Ganguly for many helpful discussions. Mr. Weihong Zhang carried out the simulations of CBP/ACTR in various GBSW and GBSW/MS2 force fields. This work was supported by an Innovative Research Award from the Terry C. Johnson Center for Basic Cancer Research and a CAREER Award from NSF (Grant MCB 0952514). This paper is Contribution No. 11-007-J from the Kansas Agricultural Experiment Station.

Supporting Information Available: Total electrostatic solvation free energies of 22 proteins in test set 1, stabilities of all 44 polar and nonpolar interactions, comparison of helicities of (AAQAA)₃ in GBSW and GBMV2, summary of (Ala)₅ backbone conformational equilibria, and CHARMM input stream files for setting up optimized GBSW/MS2 and GBMV2 implicit solvent. This information is available free of charge via the Internet at <http://pubs.acs.org/>.

References

- Ponder, J. W.; Case, D. A. *Adv. Protein Chem.* **2003**, *66*, 27–85.
- MacKerell, A. D., Jr. *J. Comput. Chem.* **2004**, *25*, 1584–1604.
- Roux, B.; Simonson, T. *Biophys. Chem.* **1999**, *78*, 1–20.
- Feig, M.; Brooks, C. L., III. *Curr. Opin. Struct. Biol.* **2004**, *14*, 217–224.
- Baker, N. A. *Curr. Opin. Struct. Biol.* **2005**, *15*, 137–143.
- Chen, J.; Brooks, C. L., III; Khandogin, J. *Curr. Opin. Struct. Biol.* **2008**, *18*, 140–148.
- Cramer, C. J.; Truhlar, D. G. *Chem. Rev.* **1999**, *99*, 2161–2200.
- Bashford, D.; Case, D. A. *Annu. Rev. Phys. Chem.* **2000**, *51*, 129–152.
- Galicchio, E.; Paris, K.; Levy, R. M. *J. Chem. Theory Comput.* **2009**, *5*, 2544–2564.
- Lee, M. S.; Olson, M. A. *J. Phys. Chem. B* **2005**, *109*, 5223–5236.
- Okur, A.; Wickstrom, L.; Simmerling, C. *J. Chem. Theory Comput.* **2008**, *4*, 488–498.
- Im, W.; Chen, J.; Brooks, C. L., III. *Adv. Protein Chem.* **2005**, *72*, 173–198.
- Chen, J.; Im, W.; Brooks, C. L., III. *J. Am. Chem. Soc.* **2006**, *128*, 3728–3736.
- Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. *Proteins* **2006**, *65*, 712–725.
- Jang, S.; Kim, E.; Pak, Y. *Proteins* **2007**, *66*, 53–60.
- Ferrara, P.; Apostolakis, J.; Caffisch, A. *Proteins* **2002**, *46*, 24–33.
- Lazaridis, T. *Proteins* **2005**, *58*, 518–527.
- Warwicker, J.; Watson, H. C. *J. Mol. Biol.* **1982**, *157*, 671–679.
- Nicholls, A.; Honig, B. *J. Comput. Chem.* **1991**, *12*, 435–445.
- Im, W.; Beglov, D.; Roux, B. *Comput. Phys. Commun.* **1998**, *111*, 59–75.
- Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. *J. Am. Chem. Soc.* **1990**, *112*, 6127–6129.
- Constanciel, R.; Contreras, R. *Theor. Chim. Acta* **1984**, *65*, 1–11.
- Levy, R. M.; Zhang, L. Y.; Galicchio, E.; Felts, A. K. *J. Am. Chem. Soc.* **2003**, *125*, 9523–9530.
- Chen, J.; Brooks, C. L., III. *J. Am. Chem. Soc.* **2007**, *129*, 2444–2445.
- Chen, J.; Brooks, C. L., III. *Phys. Chem. Chem. Phys.* **2008**, *10*, 471–481.
- Born, M. *Z. Phys.* **1920**, *1*, 45–48.
- Onufriev, A.; Case, D. A.; Bashford, D. *J. Comput. Chem.* **2002**, *23*, 1297–1304.
- Feig, M.; Onufriev, A.; Lee, M.; Im, W.; Case, D.; Brooks, C. L., III. *J. Comput. Chem.* **2004**, *25*, 265–284.
- Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem.* **1996**, *100*, 19824–19839.
- Schaefer, M.; Karplus, M. *J. Phys. Chem.* **1996**, *100*, 1578–1599.
- Qiu, D.; Shenkin, P. S.; Hollinger, F. P.; Still, W. C. *J. Phys. Chem. A* **1997**, *101*, 3005–3014.
- Scarsi, M.; Apostolakis, J.; Caffisch, A. *J. Phys. Chem. A* **1997**, *101*, 8098–8106.

- (33) Ghosh, A.; Rapp, C. S.; Friesner, R. A. *J. Phys. Chem. B* **1998**, *102*, 10983–10990.
- (34) Dominy, B. N.; Brooks, C. L., III. *J. Phys. Chem. B* **1999**, *103*, 3765–3773.
- (35) Srinivasan, J.; Trevathan, M. W.; Beroza, P.; Case, D. A. *Theor. Chem. Acc.* **1999**, *101*, 426–434.
- (36) Onufriev, A.; Bashford, D.; Case, D. A. *J. Phys. Chem. B* **2000**, *104*, 3712–3720.
- (37) Lee, M. S.; Salsbury, F. R., Jr.; Brooks, C. L., III. *J. Chem. Phys.* **2002**, *116*, 10606–10614.
- (38) Im, W.; Lee, M. S.; Brooks, C. L., III. *J. Comput. Chem.* **2003**, *24*, 1691–1702.
- (39) Gallicchio, E.; Levy, R. M. *J. Comput. Chem.* **2004**, *25*, 479–499.
- (40) Zhu, J.; Alexov, E.; Honig, B. *J. Phys. Chem. B* **2005**, *109*, 3008–3022.
- (41) Haberthur, U.; Caffisch, A. *J. Comput. Chem.* **2008**, *29*, 701–715.
- (42) MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiórkiewicz-Kuczera, J.; Yin, D.; Karplus, M. *J. Phys. Chem. B* **1998**, *102*, 3586–3616.
- (43) Feig, M.; MacKerell, A. D., Jr.; Brooks, C. L., III. *J. Phys. Chem.* **2003**, *107*, 2831–2836.
- (44) MacKerell, A., Jr.; Feig, M.; Brooks, C., III. *J. Am. Chem. Soc.* **2004**, *126*, 698–699.
- (45) MacKerell, A., Jr.; Feig, M.; Brooks, C., III. *J. Comput. Chem.* **2004**, *25*, 1400–1415.
- (46) Khandogin, J.; Chen, J.; Brooks, C. L., III. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 18546–18550.
- (47) Khandogin, J.; Raleigh, D. P.; Brooks, C. L., III. *J. Am. Chem. Soc.* **2007**, *129*, 3056–3057.
- (48) Khandogin, J.; Brooks, C. L., III. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 16880–16885.
- (49) Ganguly, D.; Chen, J. *J. Am. Chem. Soc.* **2009**, *131*, 5214–5223.
- (50) Chen, J. *J. Am. Chem. Soc.* **2009**, *131*, 2088–2089.
- (51) Im, W.; Beglov, D.; Roux, B. *Comput. Phys. Commun.* **1998**, *111*, 59–75.
- (52) Lu, Q.; Luo, R. *J. Chem. Phys.* **2003**, *119*, 11035.
- (53) Swanson, J. M. J.; Mongan, J.; McCammon, J. A. *J. Phys. Chem. B* **2005**, *109*, 14769–14772.
- (54) Lee, B.; Richards, F. M. *J. Mol. Biol.* **1971**, *55*, 379–400.
- (55) Lee, M. S.; Feig, M.; Salsbury, F. R., Jr.; Brooks, C. L., III. *J. Comput. Chem.* **2003**, *24*, 1348–1356.
- (56) Yu, Z. Y.; Jacobson, M. P.; Friesner, R. A. *J. Comput. Chem.* **2006**, *27*, 72–89.
- (57) Mongan, J.; Simmerling, C.; McCammon, J. A.; Case, D. A.; Onufriev, A. *J. Chem. Theory Comput.* **2007**, *3*, 156–169.
- (58) Chocholousova, J.; Feig, M. *J. Comput. Chem.* **2006**, *27*, 719–729.
- (59) Tjong, H.; Zhou, H. X. *J. Phys. Chem. B* **2007**, *111*, 3055–3061.
- (60) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* **1983**, *4*, 187–217.
- (61) Brooks, B. R.; Brooks, C. L.; Mackerell, A. D.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; Caffisch, A.; Caves, L.; Cui, Q.; Dinner, A. R.; Feig, M.; Fischer, S.; Gao, J.; Hodosek, M.; Im, W.; Kuczera, K.; Lazaridis, T.; Ma, J.; Ovchinnikov, V.; Paci, E.; Pastor, R. W.; Post, C. B.; Pu, J. Z.; Schaefer, M.; Tidor, B.; Venable, R. M.; Woodcock, H. L.; Wu, X.; Yang, W.; York, D. M.; Karplus, M. *J. Comput. Chem.* **2009**, *30*, 1545–1614.
- (62) Richards, F. M. *Annu. Rev. Biophys. Bioeng.* **1977**, *6*, 151–176.
- (63) Shirts, M. R.; Pande, V. S. *J. Chem. Phys.* **2003**, *119*, 5740–5761.
- (64) Shalongo, W.; Dugad, L.; Stellwagen, E. *J. Am. Chem. Soc.* **1994**, *116*, 8288–8293.
- (65) Blanco, F. J.; Rivas, G.; Serrano, L. *Nat. Struct. Biol.* **1994**, *1*, 584–590.
- (66) Fesinmeyer, R. M.; Hudson, F. M.; Andersen, N. H. *J. Am. Chem. Soc.* **2004**, *126*, 7238–7243.
- (67) Feig, M.; Karanicolas, J.; Brooks, C. L., III. *J. Mol. Graphics Modell.* **2004**, *22*, 377–395.
- (68) Lei, H.; Duan, Y. *Curr. Opin. Struct. Biol.* **2007**, *17*, 187–191.
- (69) Nymeyer, H. *J. Chem. Theory Comput.* **2008**, *4*, 626–636.
- (70) Denschlag, R.; Lingeneil, M.; Tavan, P. *Chem. Phys. Lett.* **2008**, *458*, 244–248.
- (71) Zheng, W.; Andrec, M.; Gallicchio, E.; Levy, R. M. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 15340–15345.
- (72) Periole, X.; Mark, A. E. *J. Chem. Phys.* **2007**, *126*, 014903.
- (73) Graf, J.; Nguyen, P. H.; Stock, G.; Schwalbe, H. *J. Am. Chem. Soc.* **2007**, *129*, 1179–1189.
- (74) Nina, M.; Beglov, D.; Roux, B. *J. Phys. Chem. B* **1997**, *101*, 5239–5248.
- (75) Nina, M.; Im, W.; Roux, B. *Biophys. Chem.* **1999**, *78*, 89–96.
- (76) Deng, Y.; Roux, B. *J. Phys. Chem. B* **2004**, *108*, 16567–16576.
- (77) Demarest, S. J.; Martinez-Yamout, M.; Chung, J.; Chen, H. W.; Xu, W.; Dyson, H. J.; Evans, R. M.; Wright, P. E. *Nature* **2002**, *415*, 549–553.
- (78) Ebert, M. O.; Bae, S. H.; Dyson, H. J.; Wright, P. E. *Biochemistry* **2008**, *47*, 1299–1308.
- (79) Wickstrom, L.; Okur, A.; Simmerling, C. *Biophys. J.* **2009**, *97*, 853–856.
- (80) Best, R. B.; Buchete, N.; Hummer, G. *Biophys. J.* **2008**, *95*, 4494.
- (81) Gao, Y. Q. *J. Chem. Phys.* **2008**, *128*, 064105.
- (82) Liwo, A.; Czaplowski, C.; Oldziej, S.; Scheraga, H. A. *Curr. Opin. Struct. Biol.* **2008**, *18*, 134–139.
- (83) Kang, M.; Smith, P. E. *J. Comput. Chem.* **2006**, *27*, 1477–1485.

JCTC

Journal of Chemical Theory and Computation

Exchange Often and Properly in Replica Exchange Molecular Dynamics

Daniel J. Sindhikara,[†] Daniel J. Emerson,[‡] and Adrian E. Roitberg^{*,‡}

Department of Physics, School of Science, Nagoya University, Nagoya, Aichi 464-8602, Japan, Quantum Theory Project and Department of Chemistry, University of Florida, Gainesville, Florida 32611

Received May 29, 2010

Abstract: Previous work by us showed that in replica exchange molecular dynamics, exchanges should be attempted extremely often, providing gains in efficiency and no undesired effects. Since that time some questions have been raised about the extendability of these claims to the general case. In this work, we answer this question in two ways. First, we perform a study measuring the effect of exchange attempt frequency in explicit solvent simulations including thousands of atoms. This shows, consistent with the previous assertion, that high exchange attempt frequency allows an optimal rate of exploration of configurational space. Second, we present an explanation of many theoretical and technical pitfalls when implementing replica exchange that cause “improper” exchanges resulting in erroneous data, exacerbated by high exchange attempt frequency.

1. Introduction

For molecular dynamics (MD) and Monte Carlo (MC) simulations alike, the need for sufficient sampling is fundamental. Since researchers are limited to the computational power available, there is a huge need for sampling efficiency in MD and MC algorithms. Two main factors limit sampling efficiency: phase space diffusion and kinetic trapping. Both factors are system dependent, but kinetic trapping can be avoided by utilizing an enhanced sampling technique to increase barrier-hopping. Such techniques may use a coupling to an alternate ensemble where the barrier heights are reduced. This can be done in either serial or parallel manner. Serial methods such as simulated tempering utilize estimated statistical weight factors to determine probabilities of switching between ensembles.¹ Conversely, parallel methods, e.g., parallel tempering (PT) also known as the replica exchange method, switch between ensembles using exactly known probabilities by exchanging two replicas.^{2,3} These two types of methods (including all variants thereof) utilize an MC move to change ensembles after a period of simulating the system within an ensemble with either MD or MC. It is the

choice of the simulator how long one must simulate within an ensemble before attempting the MC switch to another.

To date, two popular opinions exist about the exchange attempt frequency (EAF) between ensembles. One is that exchanges should be performed often, but no more often than the potential energy autocorrelation time.^{4,5} Periole et al.⁴ studied explicit solvent REMD simulations with EAFs of 10, 2, 0.5, and 0.2 ps⁻¹ (the MD time step was not reported in this work). They found that the 10 ps⁻¹ simulation had correlated exchanges and thus only reported convergence for the three smaller EAF simulations.⁴ Of these three, the sampling efficiency grew with EAF. Abraham et al. similarly argued against using a high EAF due to correlated exchanges, and also showed that “mixing” increased with EAF.⁵ Yang et al. argued that the sampling efficiency gained by the increased traversing of energy space using high EAF has little effect on conformational sampling and is not worth the computational cost (they used a Perl wrapper outside an MD engine to perform exchanges and hence there is a substantial computational overhead to starting and stopping the runs for the individual replicas) for their simulations of alanine-dipeptide and Ala-Pro.⁶ This argument seems to dispute the efficiency of REMD in general.

The other argument is that EAFs should be maximized to attempt exchanges every few steps. Our previous work made this argument based on an array of simulations of peptides

* Corresponding author e-mail: roitberg@qtp.ufl.edu.

[†] Department of Physics, Nagoya University.

[‡] Quantum Theory Project and Department of Chemistry, University of Florida.

in implicit solvent and a reduced analytical model.⁷ Other works have achieved identical conclusions.^{8–10} So what is the cause of the discrepancy?

Here, we present further and complementary evidence to our previous study, and show, unequivocally, that EAF should be maximized in all simulations as long as the exchanges are done correctly. In the context of REMD, we will concisely explain the conditions in which an exchange is “proper” and argue that the contradictory results presented above can be directly assigned to wrongly performed exchanges.

2. Theory

In PT, also known as temperature replica exchange molecular dynamics (T-REMD), several copies (replicas) of a system are simulated simultaneously at a ladder of temperatures. The temperatures typically range from about the one most relevant to the system to a temperature high enough to easily make pass energetic barriers. Periodically, a proposal is made to exchange conformations between replicas adjacent in temperature. The probability of exchange between a replica in conformation, i , and temperature, n , with a replica in conformation, j , at replica, m , is based on the difference in energies of the two systems as well as the difference in temperatures as shown in eq 1. The formal derivation uses the a Boltzmann weighted population as the limiting distribution after convergence, as well as imposing detailed balance between the ensembles before and after an exchange.³ Note that in this formulation, there is no mention or use for different EAFs, and should work regardless, if implemented correctly. As usual with MC algorithms, once a move is accepted, the question of the new structure belonging or not to the ensemble has a trivial answer: of course it is part of the ensemble, the positive answer to the MC question guarantees it.

$$P_{\text{acc}} = \exp(-(E_j - E_i)(\beta_n - \beta_m)) \quad (1)$$

Here, E represents the total energy and $\beta = 1/k_B T$ is the inverse temperature. The E can be replaced by the potential energy V in order to increase the acceptance rate by rescaling the velocities by a factor of $\sqrt{T_{\text{new}}/T_{\text{old}}}$ to match the new temperature in the case of an accepted exchange.³ This step is crucial: it analytically cancels out the kinetic energy terms and makes the exchange in terms of potential energy alone correct. Note that some implementations use velocity reassignment rather than rescaling.^{11,12} Velocity reassignment, however, scrambles all pre-exchange momentum and may reduce the rate of barrier crossing. Nadler et al. showed an increased acceptance rate using amplified rescaling factors.¹³ After the simulation is complete, time series data at a temperature (or temperatures) of interest are collected, or a reweighting technique such as WHAM^{14,15} or MBAR¹⁶ is used to collect data from all temperatures.

There are many factors in REMD that have previously been scrutinized. For example, the ideal distribution temperatures at which replicas are held. Most commonly, the temperatures are distributed exponentially.¹⁷ This distribution provides a uniform acceptance ratio in the limit of constant

heat capacity. Trebst et al. described a technique to further optimize this distribution of temperatures.¹⁸ The lowest temperature is usually near the temperature of highest interest. Zheng et al. showed that the maximum temperature should be just above the temperature where the activation free energy vanishes and any higher replicas than that would decrease the efficiency due to non-Arrhenius behavior.¹⁹ Nymeyer recently proved this result analytically.²⁰

Rosta et al. showed that use of thermostats that do not sample the proper canonical space (such as the Berendsen thermostat) leads to incorrect results.²¹ We also showed recently that poor implementation of REMD with respect to pseudorandom number seeds can cause negative results.²² Recent work in systems with sharp phase transitions indicate that one should use different EAF for different replicas, according to the canonical correlation time at that temperature.²³

Our previous work showed that larger EAFs are better than smaller ones.⁷ Though the result was only for implicit solvent systems, we believe the result is general and fundamental. Simply put, attempting replica exchanges more often allows for more (computationally inexpensive) MC moves; in MC simulations, it is obviously better to have performed more MC moves than less. Thus, as long as the move is thermodynamically correct and computationally inexpensive, it is worth doing.

In our previous work on EAF, we also highlighted the need for computationally efficient exchange attempt calculations.⁷ Efficient exchanging can be achieved by performing exchanges within the MD engine rather than continuous calls of the MD engine from an external wrapper. The computational costliness of an exchange attempt should be the chief concern for simulators as it is the only potential limiting factor for EAF.

3. Methods

For this work, two polyalanine peptides (Ace-A_{*n*}-Nme, for $n = 3, 7$) were simulated using REMD. Each peptide was capped by an acetyl group (Ace) at the *N*-terminus and an *N*-methylamine (Nme) group at the *C*-terminus. All simulations were performed using AMBER 10 molecular simulation package with the AMBER ff99SB parameter set.^{24,25} The SHAKE algorithm in which all bonds involving hydrogen were constrained giving a 2 fs time step.²⁶ Calculations were performed in a canonical ensemble with the Langevin thermostat (collision frequency of 1 ps⁻¹). Each molecule was solvated in an octahedral box of TIP3P water molecules. A total of 1301 water molecules were added to (Ala)₇ (total system size 3985 atoms) and 761 water molecules were added to (Ala)₃ (total size of 2325 atoms). Each system was run with REMD with temperatures exponentially distributed with a theoretical acceptance probability of 20%. For (Ala)₃, Twenty-four temperatures were used ranging from 293.2 to 495.5 K; for (Ala)₇, and twenty-eight temperatures were used ranging from 294.9 to 467.9 K. For each system, 10 ns-long test runs of varying EAFs were compared to a reference 100 ns-long REMD run. The test runs utilized an array of EAFs (Ala₃: 0.01, 0.02, 0.2, 2, 25, 62.5, and 250 ps⁻¹; Ala₇: 0.001, 0.01, 0.1, 2, 20, and 250 ps⁻¹). Since the time step is 2 fs,

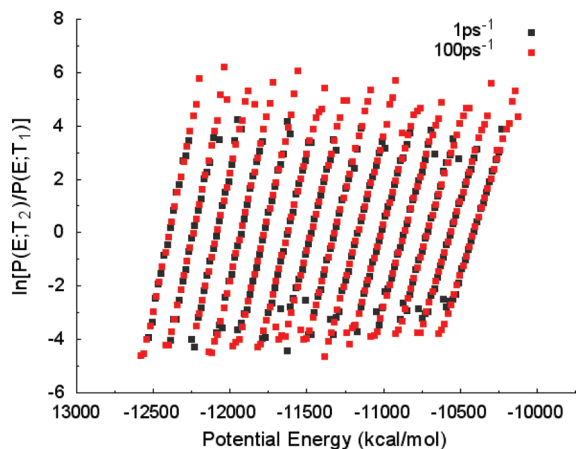


Figure 1. Logarithm of ratio of population between adjacent temperatures for simulations with different EAFs. Each of the 14 lines visible represents the overlap between two adjacent temperatures (of the 28 total, only every other neighboring pair is shown). Since the points from each EAF lie on the same line, the overlap region between the two temperature distributions must be identical.

the fastest exchange for both Ala₃ and Ala₇ was every 2 MD steps. The reference run utilized a “moderate” EAF of 0.1 ps⁻¹ (every 5000 MD steps). Each test run was compared to the reference run to judge its convergence to the “correct” value.

4. Results and Discussion

Much of the analysis following is similar to our previous work⁷ on this new data set. Here we will highlight the most important findings.

4.1. Thermal Equilibration at High EAF. To check if EAF affects thermal equilibration, energetic properties were compared between a moderate EAF of 1 ps⁻¹ and a high EAF of 100 ps⁻¹ for the (Ala)₇ simulations. The overlap in the potential energy distributions was calculated to check the similarity. For all 28 temperatures, the lowest overlap between distributions at the two EAFs is 0.996 indicating that the distributions are nearly identical. Figure 1 shows the logarithm of the ratio of energy distributions between neighboring temperatures for the (Ala)₇ simulations at both 1 ps⁻¹ and 100 ps⁻¹ (only every other pair of temperatures is shown). Since the points for either simulation lie along the same line, the two simulations must be sampling the same Boltzmann distribution.

4.2. Effect of EAF on Sampling Efficiency. Figure 2 shows the average rmsd of the dihedral population for both (Ala)₃ and (Ala)₇. This rmsd represents the difference between the test and the reference simulations, computed by constructing 36 × 36 histograms for each residue into 10° × 10° bins. These histograms were normalized into populations and the rmsd between the histograms from test and reference runs was calculated and later averaged over all residues in each peptide we studied.

Consistent with the results from the implicit solvent study,⁷ the trend is that for the larger system, the error asymptotically decreases with EAF. While there is no clear effect of EAF on efficiency for the Ala₃ simulations, mostly due to the fact

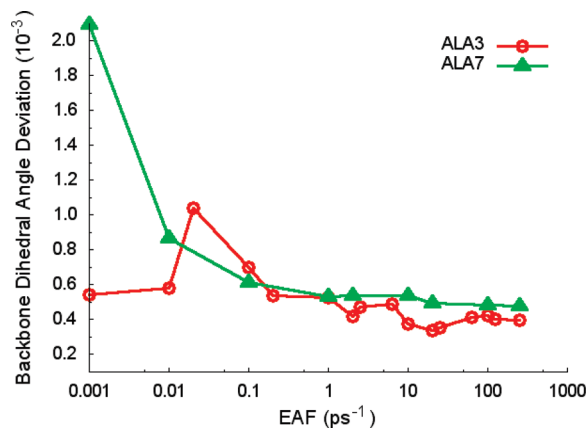


Figure 2. Backbone dihedral angle rmsd between 10 ns-long test runs and 100 ns long reference run.

that it is a very simple system to converge, the error in the Ala₇ simulations clearly decays with EAF.

Figure 3a,b shows linear-log plots of the deviation vs time for the (Ala)₃ and (Ala)₇ simulations respectively. While there is some line crossing, the errors in the simulations with the largest EAFs tend to reduce the fastest. The line crossing also indicates that preliminary runs to optimize EAF may not suggest the correct asymptotic behavior. For example, in the (Ala)₇ simulation, the simulation with the smallest EAF (0.001 ps⁻¹) seems to be on par with the rest after only 0.5 ns of simulation; eventually, however, it becomes clear that it is the least efficient of all the simulations.

Another indicator of sampling efficiency in REMD is the total number of times a replica visits both the lowest and the highest temperature during the 10 ns simulations.^{18,19,27–30} This “round trip number” for a fixed simulation time represents the speed of diffusion of a replica in temperature space and is thus indicative of the freedom of the simulation to sample. Figure 4 shows the round trip number for both the (Ala)₃ and (Ala)₇ simulations. The trend is clearly that the higher EAF simulations move faster through temperature space. It may be of interest to note that the dependence of number of round trips on EAF is logarithmic rather than linear as would be expected with freely diffusing systems. This may be caused partially by energy autocorrelation times as well as other factors that cause nonlinearity even in slow EAF simulations as clearly seen in our previous study.⁷

4.3. 1-Dimensional REMC System. Finally, to present a simple system that has the features of a larger, more complex set, we constructed a 1D test system using replica exchange Monte Carlo (REMC). In this system, it is easy to test the fundamental effect of EAF on replica exchange simulations. The energy is a simple 1D double well as in eq 2.

$$U(x) = x^2 - |x| \quad (2)$$

Two temperatures were used (300 K and 350 K) using metropolis Monte Carlo with a MC step evenly distributed within [−0.15,+0.15]. A total of 1 000 000 MC steps were taken at each temperature and replica exchanges were attempted at various EAFs ranging from 0.0001 to 1 per MC step (one exchange attempted every 10 000 MC steps, or one exchange attempted at every single MC step). The error

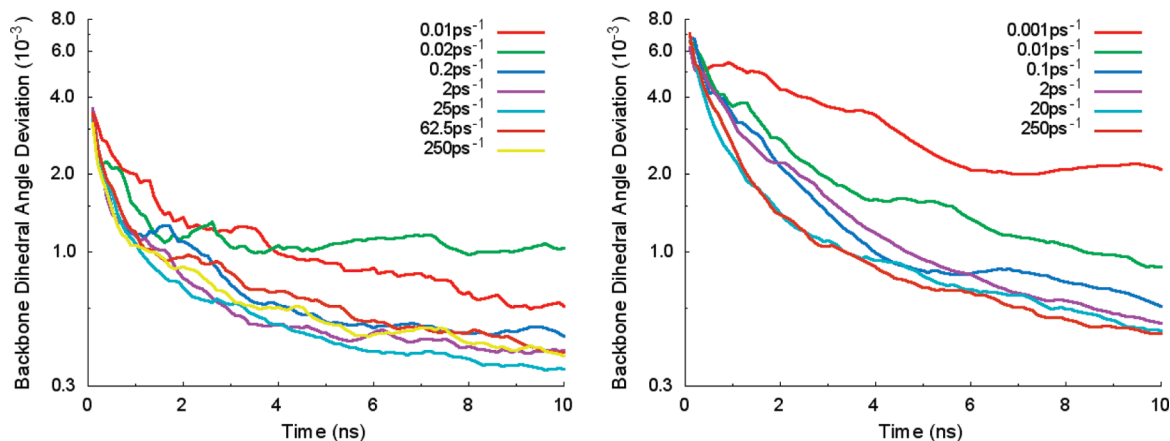


Figure 3. (a,b) Linear-log plot of deviation vs time for various EAFs. (a) (Ala)₃. (b) (Ala)₇.

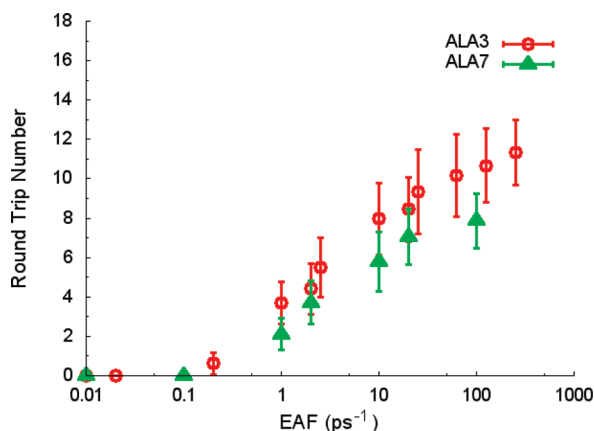


Figure 4. Average number of round trips per replica between the lowest and highest temperature for both the (Ala)₃ and (Ala)₇ simulations within 10 ns. The error is calculated as the standard deviation between individual replica counts.

Table 1. Population Error Vs Correct Distribution^a

EAF (per MC step)	0.0001	0.001	0.01	0.1	1
population error (10 ⁻³)	677	590	255	39	8

^a Error is RMSD between bin populations. The bin width is 0.1.

versus the correct, analytical solution is shown in Table 1. The trend is the same as with the molecular simulations in this and previous studies, clearly supporting the idea that larger EAFs are better than small ones.

None of these results should be surprising considering the fact that replica exchanges are not fundamentally different than any other MC move. Just as in MC, the more moves, the better, such is the case of replica exchange attempts.

4.4. Recommendations. Considering the many critiques of REMD mentioned in Section 2, along with the complementary evidence above, we emphasize that the most proper way to perform replica exchanges is as follows:

- Use a canonical thermostat (such as Langevin, Andersen, or Nose-Hoover).²¹
- Utilize a reasonable temperature distribution (as described in the Theory section).^{18–20}
- Rescale velocities after the exchange if using potential energy in the MC exchange criterion.³
- Utilize an exchange algorithm within the MD engine.
- Attempts exchanges as often as possible.*

We note that the simulator should always question the computational efficiency of an exchange in their particular algorithm and hardware. Algorithms with computationally expensive exchanges (such as those performed in an external wrapper) may not afford frequent exchanges and should be upgraded. The computationally efficient exchanges in AMBER10 (and above) allow for high EAFs and thus we found that exchanging every 10–100 steps is a good balance between sampling and computational efficiency. Furthermore, if any of the criteria above are violated, the use of high EAF could exacerbate the respective detrimental effects (as described above and within the cited references).

5. Conclusions

Despite strong evidence in prior studies that REMD using high EAFs are optimal for fastest sampling, questions remained about the consistency of this result for explicit solvent systems. Here, a similar study to the original, implicit solvent study⁷ shows that the trend also holds for explicit solvent systems. This new work should resolve many remaining questions about the effect of EAF on REMD simulations. Combined with the extensive analysis of the study, we strongly conclude that for any system where REMD exchanges are performed properly, as described above, maximum sampling efficiency can be obtained by attempting to exchange as often as possible. We also detailed many of the ways REMD simulations were performed improperly.

Acknowledgment. The authors acknowledge the University of Florida High-Performance Computing Center and NSF Large Allocations Resource Committee through Grant Nos. TG-MCA05S010 and UT-NTNL0002 for providing computational resources and support that have contributed to the research results reported within this work. Work at University of Florida was funded by the National Science Foundation Grant No. CHE-0822-935. The authors wish to thank Kevin Hauser for his initial work on these systems.

References

- (1) Marinari, E.; Parisi, G. Simulated tempering—A new Monte-Carlo scheme. *Europhys. Lett.* **1992**, *19* (6), 451–458.

- (2) Hansmann, U. H. E. Parallel tempering algorithm for conformational studies of biological molecules. *Chem. Phys. Lett.* **1997**, *281* (1–3), 140–150.
- (3) Sugita, Y.; Okamoto, Y. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* **1999**, *314* (1–2), 141–151.
- (4) Periole, X.; Mark, A. E. Convergence and sampling efficiency in replica exchange simulations of peptide folding in explicit solvent. *J. Chem. Phys.* **2007**, *126* (1), 014903.
- (5) Abraham, M. J.; Gready, J. E. Ensuring mixing efficiency of replica-exchange molecular dynamics simulations. *J. Chem. Theory Comput.* **2008**, *4* (7), 1119–1128.
- (6) Yang, L. J.; Shao, Q.; Gao, Y. Q. Comparison between integrated and parallel tempering methods in enhanced sampling simulations. *J. Chem. Phys.* **2009**, *130* (12), 124111.
- (7) Sindhikara, D.; Meng, Y. L.; Roitberg, A. E. Exchange frequency in replica exchange molecular dynamics. *J. Chem. Phys.* **2008**, *128* (2), 024103.
- (8) Opps, S. B.; Schofield, J. Extended state-space Monte Carlo methods. *Phys. Rev. E* **2001**, *6305* (5), 056701.
- (9) Zhang, C.; Ma, J. P. Comparison of sampling efficiency between simulated tempering and replica exchange. *J. Chem. Phys.* **2008**, *129* (13), 134112.
- (10) Rosta, E.; Hummer, G. Error and efficiency of replica exchange molecular dynamics simulations. *J. Chem. Phys.* **2009**, *131* (16), 165102.
- (11) Chodera, J. D.; Swope, W. C.; Pitera, J. W.; Seok, C.; Dill, K. A. Use of the weighted histogram analysis method for the analysis of simulated and parallel tempering simulations. *J. Chem. Theory Comput.* **2007**, *3* (1), 26–41.
- (12) Yang, L. J.; Grubb, M. P.; Gao, Y. Q. Application of the accelerated molecular dynamics simulations to the folding of a small protein. *J. Chem. Phys.* **2007**, *126* (12), 125102.
- (13) Nadler, W.; Hansmann, U. H. E. Optimizing replica exchange moves for molecular dynamics. *Phys. Rev. E* **2007**, *76* (5), 057102.
- (14) Ferrenberg, A. M.; Swendsen, R. H. Optimized Monte-Carlo Data-Analysis. *Phys. Rev. Lett.* **1989**, *63* (12), 1195–1198.
- (15) Kumar, S.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A.; Rosenberg, J. M. The weighted histogram analysis method for free-energy calculations on biomolecules 0.I. The method. *J. Comput. Chem.* **1992**, *13* (8), 1011–1021.
- (16) Shirts, M. R.; Chodera, J. D. Statistically optimal analysis of samples from multiple equilibrium states. *J. Chem. Phys.* **2008**, *129* (12), 124105.
- (17) Kofke, D. A. On the acceptance probability of replica-exchange Monte Carlo trials. *J. Chem. Phys.* **2002**, *117* (15), 69116914.
- (18) Trebst, S.; Troyer, M.; Hansmann, U. H. E. Optimized parallel tempering simulations of proteins. *J. Chem. Phys.* **2006**, *124* (17), 174903.
- (19) Zheng, W. H.; Andrec, M.; Gallicchio, E.; Levy, R. M. Simulating replica exchange simulations of protein folding with a kinetic network model. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104* (39), 15340–15345.
- (20) Nymeyer, H. How efficient is replica exchange molecular dynamics? An analytic approach. *J. Chem. Theory Comput.* **2008**, *4* (4), 626–636.
- (21) Rosta, E.; Buchete, N. V.; Hummer, G. Thermostat artifacts in replica exchange molecular dynamics simulations. *J. Chem. Theory Comput.* **2009**, *5* (5), 1393–1399.
- (22) Sindhikara, D. J.; Kim, S.; Voter, A. F.; Roitberg, A. E. Bad seeds sprout perilous dynamics: Stochastic thermostat induced trajectory synchronization in biomolecules. *J. Chem. Theory Comput.* **2009**, *5* (6), 1624–1631.
- (23) Bittner, E.; Nussbaumer, A.; Janke, W. Make life simple: Unleash the full power of the parallel tempering algorithm. *Phys. Rev. Lett.* **2008**, *101* (13), 130603.
- (24) Case, D. A.; Cheatham, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. The Amber biomolecular simulation programs. *J. Comput. Chem.* **2005**, *26* (16), 1668–1688.
- (25) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. Comparison of multiple amber force fields and development of improved protein backbone parameters. *Proteins: Struct., Funct., Bioinf.* **2006**, *65* (3), 712–725.
- (26) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of *N*-alkanes. *J. Comput. Phys.* **1977**, *23*, 327–341.
- (27) Fenwick, M. K.; Escobedo, F. A. Expanded ensemble and replica exchange methods for simulation of protein-like systems. *J. Chem. Phys.* **2003**, *119* (22), 11998–12010.
- (28) Rathore, N.; Chopra, M.; de Pablo, J. J. Optimal allocation of replicas in parallel tempering simulations. *J. Chem. Phys.* **2005**, *122* (2), 024111.
- (29) Earl, D. J.; Deem, M. W. Parallel tempering: Theory, applications, and new perspectives. *Phys. Chem. Chem. Phys.* **2005**, *7* (23), 3910–3916.
- (30) Katzgraber, H. G.; Trebst, S.; Huse, D. A.; Troyer, M. Feedback-optimized parallel tempering Monte Carlo. *J. Stat. Mech.* **2006**, *2006*, P03018.

CT100281C

Calculation of One- and Two-Photon Absorption Spectra of Thiolated Gold Nanoclusters using Time-Dependent Density Functional Theory

Paul N. Day,^{*,†,‡} Kiet A. Nguyen,^{†,§} and Ruth Pachter^{*,†}

Materials and Manufacturing Directorate, Air Force Research Laboratory, Wright Patterson Air Force Base, Ohio 45433, General Dynamics Information Technology, Inc., 5200 Springfield Street, Dayton, Ohio 45431, and UES, Inc., 4401 Dayton Xenia Road, Dayton, Ohio 45432

Received March 16, 2010

Abstract: The one- (OPA) and two-photon (TPA) absorption spectra have been calculated for a gold dimer, for a monothiolated gold dimer anion, for a thiolated gold cluster $[\text{Au}_{25}(\text{SH})_{18}]^{-1}$, whose structure has been determined, and for a proposed cluster $[\text{Au}_{12}(\text{SR})_9]^{+1}$ using time-dependent density functional theory (TDDFT). Geometry optimization with different exchange–correlation (X-C) functionals yielded small differences which had significant consequences in the spectra calculations. The calculated excitation energies of $\text{Au}_{25}(\text{SH})_{18}^{-1}$ are in excellent agreement with experiment when the local density approximation X α -optimized geometry is used with the B3LYP X-C functional. The CAMB3LYP and mCAM functionals yielded OPA results in good agreement with experiment for the dimer systems and the larger clusters. The super-atom theory was useful in analyzing the electronic transitions in the larger clusters. TPA was dominated by resonance effects, and the calculated cross-sections displayed a strong X-C functional dependence.

I. Introduction

Gold clusters that are smaller than about 1 nm show quantum effects,^{1–3} and clusters with 25 gold atoms have recently been reported to exhibit large two-photon absorption (TPA)^{2,4} and to be capable of serving as a platform for ligand exchange with chiral ligands.⁵ The structure of the $[\text{Au}_{25}(\text{SR})_{18}]^{-1}$ complex was successfully predicted theoretically by Akola et al.⁶ for R = CH₃ and determined experimentally by Heaven et al.⁷ and by Zhu et al.³ for R = CH₂CH₂Ph. The structure, shown in Figure 1 with R = H, has an Au₁₃ icosahedral core with one Au atom at each vertex of the icosahedron and one at the center. This icosahedral core is surrounded by three pairs of V-shaped structures, (–SR–Au–SR–Au–SR–). These three pairs of Vs are in orthogonal planes which are mirror planes in the icosahedral core, reducing the effective symmetry of the

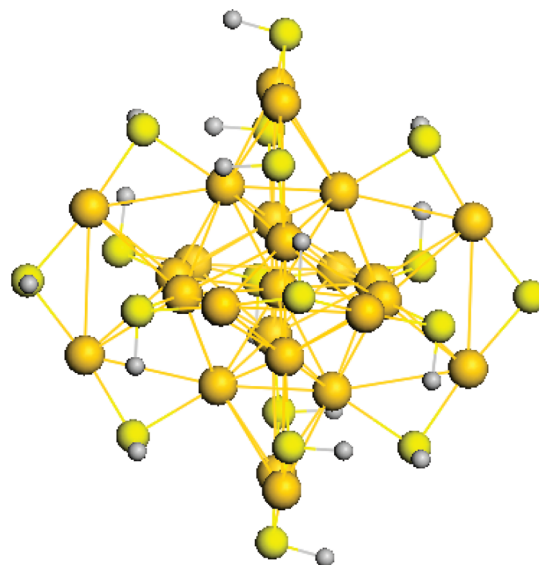


Figure 1. Optimized structure of $\text{Au}_{25}(\text{SH})_{18}^{-1}$.

molecule to D_{2h} . Each vertex atom in the icosahedral core is bonded to a sulfur atom. Of the 20 triangular faces in the

* Corresponding authors. E-mail: Paul.Day@wpafb.af.mil (P.N.D.) or Ruth.Pachter@wpafb.af.mil (R.P.).

† Wright Patterson Air Force Base.

‡ General Dynamics Information Technology, Inc.

§ UES, Inc.

icosahedral core, 12 are covered by an outer gold atom, while the other 8 faces appear unprotected.

A theory for magic number stabilities of gold clusters has been proposed by Walter et al.⁸ based in part on an earlier super-atom theory which produced magic numbers for sodium atom clusters, as shown by Knight et al.⁹ and theorized by Ekardt.^{10,11} The magic numbers occur when the number of valence electrons in the super-atom corresponds to a strong electron shell closure. Analogous with the atomic s and p orbitals, filling the 1S super-atom shell gives a magic number of 2 and filling the 1S and 1P shells gives a magic number of 8. The number of valence electrons n^* is given by

$$n^* = Nv_A - M - z$$

where N is the number of gold atoms, v_A is the valence of the gold atom (assumed to be 1), M is the number of electron-withdrawing ligands, and z is the net charge of the molecule. Thus, for $[\text{Au}_{25}(\text{SR})_{18}]^{-1}$, n^* is equal to 8, corresponding to the super-atom electron occupation $1S^21P^6$, while for $\text{Au}_{102}(\text{SR})_{44}$, for which the structure has been determined by Jadzinsky et al.,¹ n^* is equal to 58, corresponding to the occupation $1S^21P^61D^{10}2S^21F^{14}2P^61G^{18}$. This formula was exploited by Jiang and Dai¹² to design molecules based on $[\text{Au}_{25}(\text{SR})_{18}]^{-1}$ but with a different central atom; they adjusted the charge z in order to maintain $n^* = 8$. Since 2 is also a magic number, both the gold dimer and its simplest thiolated anionic structure, $[\text{Au}_2(\text{SR})]^{-1}$, should be stable. The gold dimer has been studied previously both experimentally^{13–19} and theoretically,^{20–23} but to our knowledge $[\text{Au}_2(\text{SR})]^{-1}$ has not been previously studied. We have therefore investigated these systems as well as a previously proposed²⁴ structure for the smallest thiolated gold super-atom complex, $\text{Au}_{12}(\text{SR})_9^{+1}$, to test the effect of the sulfur ligands on the absorption spectrum of very small gold clusters.

Overall, theoretical studies^{25,26} have indicated only a weak dependence of structure, energetics, and spectra on the size of the alkyl group in gold alkyl thiolates as well as on the solvent environment. The proposed reason for the small effects of solvent and ligand size on the absorption spectrum of $[\text{Au}_{25}(\text{SR})_{18}]^{-1}$ is that the relevant transitions are primarily in the gold core. The linear absorption spectrum for $[\text{Au}_{25}(\text{SR})_{18}]^{-1}$ has been measured in water with SR = glutathione,²⁷ in toluene²⁸ and hexane^{2,4} with R = *n*-hexyl, and in toluene with R = ethyl phenyl,³ and they are all in good agreement, again indicating a weak dependence on the ligand size and the solvent. Thus, in this study, for computational economy, we will focus on the smallest possible ligand with R = H for the thiolated Au_{25} cluster, and we have not included any solvent effects in the calculations.

Although time-dependent density functional theory (TDDFT) has been used to calculate the spectra for gold clusters as large as 146 gold atoms²⁹ as well as for $[\text{Au}_{25}(\text{SR})_{18}]^{-1}$ using several functionals,²⁶ this work is the first to include hybrid functionals to determine the dependence of the calculated spectra on the choice of X-C functional. In this work, we study the absorption spectrum for $[\text{Au}_{25}(\text{SH})_{18}]^{-1}$, in comparison to experiment^{2–4,27,28} and previous calculations.^{3,26} Moreover, we extend the calcula-

tions to explain the large TPA cross-sections that have been reported from experiment.^{2,4}

II. Theory

The expression for calculating the TPA cross-section has been given previously^{30–34}

$$\delta_{f0}^I(E_1 + E_2) = \frac{8\pi^4}{(\text{ch})^2} E_1 E_2 g(E_1 + E_2) |S_{f0}(u_1, u_2)|^2 \quad (1)$$

where g is a line width function, $|S_{f0}(u_1, u_2)|^2$ is the two-photon probability corresponding to a transition from the ground (0) to a final state (f),^{35–43}

$$|S_{f0}(u_1, u_2)|^2 = \left| \sum_i^N \left[\frac{(\mathbf{u}_1 \cdot \boldsymbol{\mu}_{i0})(\boldsymbol{\mu}_{if} \cdot \mathbf{u}_2)}{E_i - E_1 + i\Gamma_i} + \frac{(\mathbf{u}_2 \cdot \boldsymbol{\mu}_{i0})(\boldsymbol{\mu}_{if} \cdot \mathbf{u}_1)}{E_i - E_2 + i\Gamma_i} \right] \right|_2^2 \quad (2)$$

and E_1 and E_2 are the energies of the two photons with unit polarization vectors \mathbf{u}_1 and \mathbf{u}_2 , respectively. The transition dipole moments are given by μ_{ij} , the state energies by E_i , and the state decay constants by Γ_i . In molecules which are close to being centrosymmetric, the TPA cross-section can be estimated by the three-state approximation, which assumes a single term dominates in the sum-over-states (SOS) in eq 2, and in the case where the two photons have the same energy (E_λ), the cross-section is given by

$$\delta_{f0}^I = \frac{32\pi^4 g_{\text{max}} E_\lambda^2}{15(\text{ch})^2 (E_i - E_\lambda)^2} |\mu_{0i}|^2 |\mu_{if}|^2 (2\cos^2 \Theta_{\mu\mu} + 1) \quad (3)$$

where g_{max} is the maximum of the line width function, and $\Theta_{\mu\mu}$ is the angle between the two transition dipole moment vectors. While the three-state approximation is often not adequate for quantitative results, it can be useful in analyzing the origin of the TPA intensity, as we have previously shown.³⁴

TPA has been calculated by the above formalism for a number of organic systems,^{30,33,34,44–48} which generally only have a few states near the visible spectrum, and thus the SOS can be rapidly converged. The decay constant Γ_i in eq 2 makes a significant contribution only near resonance, i.e., when the photon energy is very close to a one-photon allowed state energy. In gold clusters, the density of states is much higher, and a near resonant condition is highly likely and generally dominates the TPA. This makes the inclusion of an appropriate decay constant in eq 2 more important in the calculation of TPA in gold clusters than in organic chromophores. The SOS method enables introduction of few-state models, which can be used with the calculation of transition dipole moments between excited states from the double residue of the response function. As described in the next session, the most computationally efficient method for calculation of the TPA is through the quadratic response single residue, and thus this method must be used for the larger clusters. However, in the single residue methodology the TPA probability is obtained directly without carrying out a SOS, so the damping factor is not a part of the formalism. However, a qualitative understanding was provided. Al-

though inorganic and hybrid structures have been previously applied in nonlinear optical applications,⁴⁹ the use of gold thiolate cluster compounds of high stability provides a platform for gaining a more fundamental understanding.

III. Computational Methods

The importance of including relativistic effects in calculations on gold has been well established.⁵⁰ In order to avoid the computational complexity of including relativistic effects through solution of the Dirac equation, approximations have been developed, such as the Douglas–Kroll (DK)^{51,52} method, which is tested here in the GAMESS⁵³ and Dalton^{54,55} programs, and the zero-order regular approximation (ZORA),^{56,57} which is used in the ADF^{58–60} program. The most computationally efficient method to include relativistic effects is through pseudopotentials. When using pseudopotentials, so-called effective core potentials (ECP), the core electrons are removed from the full calculation, and their effects are included through a simpler scaled method, greatly reducing the amount of computation. In addition, the relativistic effects can be included in the parametrization at almost no additional cost. Well parametrized pseudopotentials can be competitive in accuracy with the more expensive all-electron DK and ZORA methods. In order to check the effects of spin–orbit coupling, the two following methods were tested: the perturbative inclusion of spin–orbit coupling (PSOC)⁶¹ and the relativistic spin–orbit ZORA (RSOZ)⁶² (available in ADF).

For the isolated gold dimer, the experimental bond length of 2.472 Å¹⁷ has been used in the TDDFT calculations. The molecular geometry for [Au₂(SH)]⁻¹ was optimized with the BP86^{63,64} X-C functional and the improved model core potential with scaled relativistic effects (iMCP-SR2),⁶⁵ hereafter abbreviated SR2, and the corresponding basis set in the GAMESS⁵³ program. The geometry of [Au₂₅(SR)₁₈]⁻¹ for R = H was initially optimized with the ADF program at the BP86/ZORA/TZP level of theory with a frozen core of 4f for gold and 2p for sulfur. A recent study²⁶ indicated better agreement with the experimental structure when the exchange-only X α functional was used in the geometry optimization, and therefore a second optimized geometry was obtained for this molecule using this functional, and both geometries were tested in TDDFT calculations. The X α functional is the Slater⁶⁶ exchange functional with $\alpha = 0.7$ and is equivalent to local density approximation (LDA) if correlation is ignored.

For use with the DK method, a basis set for gold has been developed from the general contractions of Tsuchiya et al.^{67,68} and labeled “b41”. This basis set is a 26s23p15d10f/11s10p7d3f contraction and is given in Table S1 of the Supporting Information. Linear response TDDFT calculations were carried out with this basis set and the DK method, with DZP and TZP basis sets and the ZORA method, and with the SR2, LANL2DZ,⁶⁹ and SDD-DZ^{70,71} pseudopotentials and corresponding basis sets. Since quadratic response TDDFT is available only in the Dalton⁵⁵ program, the TPA results are limited to the basis sets b41/DK, LANL2DZ, and SDD. Linear response TDDFT calculations were carried out with the generalized gradient approximation (GGA) func-

tional BP86, the hybrid functional B3LYP,^{63,72–75} and the long-range corrected (LC) functionals CAMB3LYP,⁷⁶ mCAM,^{33,48} and SAOP,⁷⁷ while the quadratic response calculations were limited to the X-C functionals BP86, B3LYP, CAMB3LYP, and mCAM. For the dimer systems, OPA were also obtained from high-level coupled cluster methods, including the completely renormalized equation-of-motion coupled cluster singles and doubles with perturbative triples (CR-EOMCCSD(T))^{78–80} in GAMESS and the similar CCSDR(3)⁸¹ method in Dalton. OPA extinction coefficients and TPA cross-sections were obtained from the calculated oscillator strengths and TPA probabilities, respectively, by fitting to Gaussian line width functions, as described previously.³² The full-width at half-maximum (fwhm) used in the line width functions was 0.2 and 0.3 eV for OPA and TPA, respectively.

The TPA spectrum has been calculated with the Dalton⁵⁵ program, using both the double residue method to obtain the excited-state transition dipole moments for use in the SOS expression (eq 2) and the single residue method to directly obtain the TPA probability.^{82–85} While the single residue method is more computationally efficient, the SOS method allows for the inclusion of the intermediate-state decay constant (also called the damping constant or broadening factor, Γ_i in eq 2). In the calculation of the TPA cross-section using quadratic response TDDFT, resonance effects can produce unphysically large TPA cross-sections, and this damping constant is needed to mitigate this problem. Combining calculated TPA cross-sections with classical laser pulse propagation simulations has been proposed by Agren et al.⁸⁶

IV. Results

A. Au₂ and [Au₂SH]⁻¹. While the TDDFT calculations were carried out at the experimental geometry, geometry optimization of this smallest gold cluster was carried out with DFT to test several X-C functionals, including the hybrid B3LYP and the meta-hybrid M06.⁸⁷ The results are given in Table S2 in the Supporting Information. As was found for the Au₂₅ system, exchange-only LDA functionals, such as X α and Slater, give good agreement with the experimental geometry, while GGA X-C functionals, such as BP86, tend to overestimate the bond length, and the overestimation is even worse when the hybrid B3LYP functional is used. Including the empirical dispersion correction of Grimme⁸⁸ yields a small increase in the bond length, which slightly worsens the agreement with experiment.

The calculated OPA energies and oscillator strengths for Au₂ are given in Table 1 and compared to experiment, while the calculated values for Au₂SH⁻¹ are given in Table 2. High-level theoretical results from CCSD(T) were also used to evaluate the TDDFT methods. Experimentally, the gold dimer is characterized by the *A* and *B* lines measured at 2.44 and 3.18 eV, respectively, with corresponding reported oscillator strengths of 0.05 and 0.13.^{13,14} In most of the calculated results, the *A* line is considerably weaker than measured, while the *B* line is stronger than experiment, both of which may be due to the effects of vibronic coupling in

Table 1. OPA Excitation Energies and Oscillator Strengths for Au₂

functional	basis set	Π_u		Σ_u^+	
		$\Delta E(\text{eV})$	f	$\Delta E(\text{eV})$	f
BP86	LANL2DZ	2.480	0.007	2.909	0.160
B3LYP	LANL2DZ	2.808	0.008	3.075	0.200
CAMB3LYP	LANL2DZ	2.989	0.010	3.115	0.211
mCAMB3LYP	LANL2DZ	2.883	0.009	3.093	0.206
BP86	SDD-DZ	2.363	0.007	2.850	0.128
B3LYP	SDD-DZ	2.703	0.008	2.992	0.170
CAMB3LYP	SDD-DZ	2.882	0.009	3.016	0.180
mCAMB3LYP	SDD-DZ	2.773	0.008	3.002	0.175
BP86	IMCP-SR2	2.655	0.007	2.929	0.152
B3LYP	IMCP-SR2	2.830	0.008	3.075	0.172
BP86	B41/DK	2.311	0.007	2.813	0.129
B3LYP	B41/DK	2.650	0.009	2.976	0.169
CAMB3LYP	B41/DK	2.848	0.011	3.023	0.182
mCAMB3LYP	B41/DK	2.723	0.009	2.993	0.174
BP86	DZ.4d/ZORA	2.734	0.008	3.011	0.147
BP86	DZ.4f/ZORA	2.726	0.008	3.009	0.149
BP86	TZP.4d/ZORA	2.426	0.007	2.867	0.132
BP86	TZP.4f/ZORA	2.416	0.007	2.863	0.133
SAOP	TZP/ZORA	2.886	0.008	3.096	0.166
SAOP-PSOC	TZP/ZORA	2.877	0.005	3.273	0.128
				2.344	0.041
SAOP-RSOZ	TZP/ZORA	2.900	0.005	3.304	0.156
				2.350	0.020
CR-EOMCCSD(T)	IMCP-SR2	2.968	0.012	2.895	0.274
CCSDR(3)	LANL2DZ	2.867	0.013	2.884	0.222
CCSDR(3)	SDD-DZ	2.733	0.012	2.764	0.199
experiment ¹³		2.435	0.050	3.184	0.130

Table 2. OPA for Au₂SH⁻¹

functional	basis set	Π_u		Σ_u^+	
		$\Delta E(\text{eV})$	f	$\Delta E(\text{eV})$	f
BP86	LANL2DZ	2.7287	0.0014	3.3176	0.1200
B3LYP	LANL2DZ	3.1101	0.0022	3.4351	0.1770
CAMB3LYP	LANL2DZ	3.5330	0.0015	3.4699	0.1853
mCAMB3LYP	LANL2DZ	3.2995	0.0020	3.4562	0.1843
BP86	SDD-DZ	2.5943	0.0007	3.0831	0.1082
B3LYP	SDD-DZ	2.9702	0.0014	3.2014	0.1597
CAMB3LYP	SDD-DZ	3.3783	0.0011	3.2432	0.1689
mCAMB3LYP	SDD-DZ	3.1531	0.0013	3.2249	0.1664
BP86	B41/DK	2.8541	0.0004	3.6725	0.1098
B3LYP	B41/DK	3.3159	0.0004	3.8978	0.1483
CAMB3LYP	B41/DK	3.7555	0.0000	3.9599	0.1595
mCAMB3LYP	B41/DK	3.5029	0.0002	3.9291	0.1536
BP86	ZORA/TZP.4f	2.6785	0.0014	3.3129	0.1276
SAOP	ZORA/TZP	3.0142	0.0027	3.6610	0.2293
CR-EOMCCSD(T)	IMCP-SR2	3.4450	0.0009	3.6980	0.2745
CCSDR(3)	LANL2DZ	3.2874	0.0015	3.4366	0.1403
CCSDR(3)	SDD-DZ	3.1951	0.0010	3.3431	0.1297

the measured absorption spectra, an effect not included in the calculations. Another possible explanation is spin-orbit coupling, which will be discussed in the next paragraph. Thus, the comparison to the CCSD(T) results may be more useful for evaluation of the X-C functionals. Also, while experimental results are not available for [Au₂SH]⁻¹, the performance of the X-C functionals on this simple gold thiolate can also be evaluated by comparison to CCSD(T) results. As in a previous TDDFT study by Wang et al.,²³ the *A* line in the gold dimer was assigned to the excitation to the first Π_u state, while the stronger *B* line was assigned to the transition to the first Σ_u^+ state. The calculated *A* line, in addition to being considerably weaker than experiment,

is blue-shifted by 0.3 to 0.5 eV, while the calculated *B* line is red-shifted compared to experiment by 0.2 to 0.4 eV. As a result, in the calculated results, the two states are nearly accidentally degenerate and would not be distinguishable as separate absorption lines. For example, at the highest theoretical level carried out, CCSD(T), the difference in energy of these two final states is only about 0.02 eV (Table 1). The TDDFT results using the CAMB3LYP and mCAM functionals are in the best agreement with CCSD(T) for Au₂, yielding excitation energies with less than 0.1 eV error for the *A* line and 0.1 to 0.2 eV for the *B* line. The oscillator strengths calculated using these two functionals are only slightly smaller than the CCSD(T) values, and the calculated split between the *A* and *B* line is about 0.2 eV. The results using the B3LYP and SAOP functionals are also in good agreement with the CCSD(T) values, while for the BP86 functional the agreement is not as good. To test the possibility that the discrepancy between the calculated and measured linear absorption spectra is due to the effects of spin-orbit coupling, linear TDDFT calculations were carried out using the relativistic spin-orbit ZORA (RSOZ)⁶² and the more computationally efficient perturbative inclusion of spin-orbit coupling (PSOC).⁶¹ The results of these calculations, carried out with the SAOP X-C functional, are also included in Table 1. The difference is the existence of another Σ_u^+ state, at an excitation energy of 2.3 eV, which appears in both of these calculations. The calculated oscillator strength and excitation energy of this state is in much better agreement with the measured *A* line than the calculated Π_u state. Furthermore, the calculated oscillator strength of the transition to the Σ_u^+ state near 3.2 eV was reduced, making it in better agreement with the measured *B* line. Thus, explicit inclusion of spin-orbit coupling effects by one of these methods is recommended for Au₂. The RSOZ and PSOC methods were also tested on the Au₂₅(SH)₁₈⁻¹ and Au₁₂(SH)₉⁺¹ systems, but the effects on the calculated spectra were small. Thus, these methods are not included in the results given for these systems. The first geometry considered for [Au₂(SH)]⁻¹ was a bridged C_{2v} structure with the thiolate bonded to both gold atoms. This was found to be a saddle point, and instead the minimum-energy structure was found to have the sulfur atom bonded to just one of the gold atoms, nearly collinear with the two gold atoms. The optimized geometry has the following bond lengths in Å (compared to the experimental value for Au₂ of 2.472): Au–Au = 2.5803, Au–S = 2.3054, and S–H = 1.3538; and bond-angles in degrees (°): Au–Au–S = 179.3 and Au–S–H = 95.8. The molecule is nearly planar with the dihedral angle Au–Au–S–H equal to 173.3°. The effect of thiolation on the absorption spectrum of the gold dimer, as calculated at the CCSD(T) level, is to blue-shift the *A* line by 0.45 eV and decrease its intensity by a factor of 10 and to blue-shift the *B* line by 0.55 eV and decrease its intensity by nearly a factor of 2. The TDDFT results using the CAMB3LYP and mCAM functionals with pseudopotentials are in good agreement with the CCSD(T) results.

The TPA cross-sections calculated for the gold dimer with the different X-C functionals are not as consistent as the OPA spectra. The TPA results for Au₂ are given in Table 3. The

Table 3. TPA for Au₂^a

functional	basis set	ΔE (eV)	σ_2 (GM)	μ_{01}^2	μ_{12}^2	R	E_1 (eV)
BP86	LANL2DZ	6.200	1.500×10^4	1.45×10^1	1.36×10^2	1.03×10^3	2.91
B3LYP	LANL2DZ	6.260	1.654×10^6	1.71×10^1	9.40×10^2	1.34×10^4	3.08
CAMB3LYP	LANL2DZ	6.420	1.387×10^5	1.79×10^1	2.16×10^2	4.72×10^3	3.11
mCAMB3LYP	LANL2DZ	6.315	7.398×10^5	1.76×10^1	5.72×10^2	9.57×10^3	3.09
BP86	SDD-DZ	6.027	1.577×10^4	1.19×10^1	1.28×10^2	1.36×10^3	2.85
B3LYP	SDD-DZ	5.983	1.548×10^{13}	1.50×10^1	2.23×10^6	5.59×10^7	2.99
CAMB3LYP	SDD-DZ	6.127	8.237×10^5	1.57×10^1	4.02×10^2	1.71×10^4	3.02
mCAMB3LYP	SDD-DZ	6.032	1.118×10^8	1.54×10^1	5.44×10^3	1.74×10^5	3.00
BP86	B41/DK	7.100	2.230×10^2	1.21×10^1	2.21×10^1	9.25×10^1	2.81
B3LYP	B41/DK	7.265	4.610×10^2	1.50×10^1	3.11×10^1	1.23×10^2	2.98
CAMB3LYP	B41/DK	7.390	3.585×10^2	1.59×10^1	2.40×10^1	1.21×10^2	3.02
mCAMB3LYP	B41/DK	7.315	4.267×10^2	1.54×10^1	5.35×10^1	1.21×10^2	2.99

^a Including the transition energy (ΔE), the TPA cross-section (σ_2), the squares of the most relevant transition dipole moments (μ), the resonance enhancement factor (R), and the intermediate-state excitation energy (E_1).

calculations using the DK approximation yield the smallest and most realistic TPA cross-sections, ranging from 223 to 461 GM, while the largest cross-sections are obtained using the SDD-DZ pseudopotential, ranging from 10^4 to 10^{13} GM. While each TPA cross-section reported in Table 3 was obtained from the residue of the quadratic response function and is thus equivalent to the full SOS approximation, also listed are the transition dipole moments and resonance enhancement factors (R), which can be used in the three-state approximation. As the three-state approximation yields a good prediction for this system, these values indicate the source of the variation in the calculated TPA. When SDD is used, the two-photon state is about 6 eV above the ground state, and since the strong OPA state is near 3 eV, R is large. Since the energy of the TPA state in the DK calculations is about 7.2 eV, R is much smaller. Also, in Figure S1 in the Supporting Information, the square of each relevant excited-state transition dipole moment obtained in a quadratic response calculation is plotted against the resonance enhancement factor, and there appears to be a correlation. Methods which have a large value for the resonance enhancement factor also have a large value for the excited-state transition dipole moment, indicating a relationship in these quantities. The TPA data for $[\text{Au}_2(\text{SH})]^{-1}$ is given in Table S3 in the Supporting Information. Some of the same trends are observed, with the DK method yielding the smallest peak cross-sections (1500–475 000 GM) and the SDD pseudopotential yielding the largest (10^7 – 10^{16} GM). However, the reduction in symmetry for this compound increases the density of allowed states and opens more possibilities for highly resonant terms. As a result, the calculated peak TPA cross-section occurs over a range of transition energies from 6.68–7.05 eV for LANL2DZ, from 6.40–8.99 eV for SDD-DZ, and from 7.76–8.03 eV for the DK method. In the SOS method, the OPA decay constant broadens the intermediate state and thus dampens the TPA near the resonance, yielding a more physically realistic TPA cross-section.

B. Au₂₅(SH)₁₈⁻¹. I. OPA. In Table 4, the calculated linear absorption spectrum for Au₂₅(SH)₁₈⁻¹ is compared to the measured spectra of Negishi et al.,^{27,28} Ramakrishna et al.,^{2,4} and Zhu et al.³ Four absorption peaks, labeled a–d, were identified from the spectrum of Negishi et al.,²⁷ including absolute extinction coefficients based on the value of 8800

M⁻¹cm⁻¹ at 670 nm (1.85 eV). Zhu et al.³ identified the first three peaks in excellent agreement with Negishi et al.,²⁷ although no absolute intensities were given. The extinction coefficients in the spectrum of Ramakrishna et al.² were scaled by 0.1 to be consistent with the intensities of Negishi et al.,²⁷ and the four peaks are identified in good agreement with Negishi et al.²⁷ As was found by Aikens,²⁶ the calculated spectrum for Au₂₅(SH)₁₈⁻¹ is in best agreement with experiment when the geometry optimized with the LDA exchange-only functional X α is used. When the geometry optimized with the GGA (BP86) X-C functional was used in the linear response calculations, the average deviation from the experimental excitation energies was larger than when the geometry optimized with X α was used, and this can be explained by comparison with the experimental geometry of the Au₁₃ icosahedral core. The average distance of the icosahedral vertex Au atoms from the central Au atom is about 2.78 Å in both the experimental and X α geometries, while it is 2.85 Å in the BP86 geometry. Geometry optimization was also carried out with the X α functional and the dispersion model of Grimme, but this caused considerable distortion in the core and raised the average distance to 2.84 Å.

A number of combinations of X-C functionals and basis sets have been used in linear response calculations with these two geometries, as given in Table 4. The spectrum obtained using the B3LYP X-C functional with the SDD basis set on the X α optimized geometry gives the best agreement with experiment, while using the mCAM functional with this basis and geometry also yields excellent agreement. The results using B3LYP, mCAM, and CAMB3LYP with the BP86 optimized geometry are also in good agreement with experiment. When TDDFT is carried out using the BP86 and SAOP X-C functionals, the excitation energy for the first peak is underestimated, but agreement is fair for the other peaks.

Table 5 lists the Kohn–Sham orbital gaps for the relevant transitions and the corresponding linear response excitation energy for the SAOP, B3LYP, and mCAM functionals at the two geometries. As pointed out previously by Akola et al.⁶ and Aikens,⁸⁹ in the super-atom model this system has an occupation of $1\text{S}^21\text{P}^61\text{D}^02\text{S}^0\dots$, so the highest occupied molecular orbital (HOMO) is an approximate triply degenerate 1P (Figure S2 in the Supporting Information), while the expected quintuply degenerate 1D is split by the octahedral

Table 4. OPA Excitation Energies and Extinction Coefficient Maxima for Au₂₅(SR)₁₈⁻¹ with Various X-C Functionals, Basis Sets, and Geometries^a

TDDFT	geometry		peaks			
			a	b	c	d
measured ^{27,28}		$\Delta E(\text{eV})$	1.851	2.755	3.092	3.780
		ϵ_{max}	8.80×10^3	2.70×10^4	3.34×10^4	6.52×10^4
measured ^{2,4} (scaled by $\times 0.1$)		$\Delta E(\text{eV})$	1.834	2.878	3.166	3.727
		ϵ_{max}	9.43×10^3	3.08×10^4	3.76×10^4	5.56×10^4
measured ³ BP86/ZORA	BP86	$\Delta E(\text{eV})$	1.800	2.750	3.100	
		ϵ_{max}	8.91×10^3	4.16×10^4	3.27×10^4	5.92×10^4
SAOP/ZORA	BP86	$\Delta E(\text{eV})$	1.408	2.573	3.393	3.883
		ϵ_{max}	8.91×10^3	4.16×10^4	3.27×10^4	5.92×10^4
BP86/LANLDZ	BP86	$\Delta E(\text{eV})$	1.501	2.611	2.896	3.301
		ϵ_{max}	1.08×10^4	3.42×10^4	3.03×10^4	1.05×10^4
BP86/LANLDZ	BP86	$\Delta E(\text{eV})$	1.468	2.678	3.193	
		ϵ_{max}	9.29×10^3	4.41×10^4	7.52×10^3	
B3LYP/LANLDZ	BP86	$\Delta E(\text{eV})$	1.763	3.013	2.818	3.333
		ϵ_{max}	1.54×10^4	3.38×10^4	1.98×10^4	5.48×10^4
mCAM/LANLDZ	BP86	$\Delta E(\text{eV})$	1.881	3.056	3.311	3.626
		ϵ_{max}	1.79×10^4	3.01×10^4	3.10×10^4	7.07×10^4
BP86/SDD-DZ	BP86	$\Delta E(\text{eV})$	1.448	2.613	3.278	
		ϵ_{max}	9.24×10^3	4.04×10^4	9.33×10^3	
B3LYP/SDD-DZ	BP86	$\Delta E(\text{eV})$	1.754	2.949	2.819	3.274
		ϵ_{max}	1.55×10^4	3.09×10^4	2.64×10^4	5.28×10^4
mCAM/SDD-DZ	BP86	$\Delta E(\text{eV})$	1.876	2.991	3.29	3.561
		ϵ_{max}	1.79×10^4	3.05×10^4	2.92×10^4	6.87×10^4
CAMB3LYP/SDD-DZ	BP86	$\Delta E(\text{eV})$	2.048	2.863	3.238	3.948
		ϵ_{max}	2.08×10^4	3.50×10^3	4.39×10^4	9.73×10^4
SAOP/ZORA	X α	$\Delta E(\text{eV})$	1.602	2.592	2.937	3.522
		ϵ_{max}	1.14×10^4	2.82×10^4	2.38×10^4	2.68×10^4
B3LYP/LANLDZ	X α	$\Delta E(\text{eV})$	1.861	2.881	2.956	3.381
		ϵ_{max}	1.58×10^4	2.60×10^4	2.52×10^4	4.74×10^4
B3LYP/SDD-DZ	X α	$\Delta E(\text{eV})$	1.844	2.754	2.954	3.324
		ϵ_{max}	1.57×10^4	2.47×10^4	2.02×10^4	4.38×10^4
mCAM/SDD-DZ	X α	$\Delta E(\text{eV})$	1.958	2.928	3.258	3.593
		ϵ_{max}	1.81×10^4	3.35×10^4	1.92×10^4	5.94×10^4

^a Geometry optimizations were carried out with the ZORA/TZP basis set.**Table 5.** Calculated Kohn–Sham Orbital Differences (eV) and Corresponding Peaks in the TDDFT Linear Response Spectrum of Au₂₅(SR)₁₈^{-1a}

TDDFT	geometry		1P \rightarrow 1D _{zz}		1P \rightarrow 1D _{xy}		1P \rightarrow 2S	
			H \rightarrow L	H-2 \rightarrow L	H \rightarrow L+1	H \rightarrow L+2	H-5 \rightarrow L	
SAOP/TZP ³	BP86/TZP	K–S gap (eV)	1.37	2.42	2.31	2.52		
		linear response	1.52		2.63		2.91	
SAOP/TZP ⁹⁴	X α /TZP	K–S gap (eV)	1.48	2.39	2.30	2.73 ^b		
		linear response	1.63	2.48	2.59	2.77 ^b	2.97	
SAOP/TZP	BP86/TZP	K–S gap (eV)	1.34	2.42	2.31	2.43	2.81	
		linear response	1.50	2.48	2.61	2.52	2.90	
SAOP/TZP	X α /TZP	K–S gap (eV)	1.46	2.42	2.31	2.64 ^b	2.86	
		linear response	1.60	2.49	2.59	2.63 ^b	2.94	
B3LYP/SDD	BP86/TZP	K–S gap (eV)	2.24	3.53	3.23		3.82	
		linear response	1.75	2.95	2.82	2.77	3.27	
B3LYP/SDD	X α /TZP	K–S gap (eV)	2.34	3.55	3.20		3.87	
		linear response	1.84	2.95	2.75	3.04	3.32	
mCAM/SDD	X α /TZP	K–S gap (eV)	3.05	4.39	3.97		4.72	
		linear response	1.96	3.26	2.93	2.33	3.59	
		expt ³	1.80		2.75		3.10	

^a The experimental peaks are also listed for comparison. ^b The final state is L+3 in this calculation.

field of the six ligand wings into an approximately doubly degenerate lowest unoccupied molecular orbital (LUMO) (D_z^2 , $D_x^2 - y^2$; Figure S3 in the Supporting Information) and triply degenerate LUMO+1 (D_{xy} , D_{yz} , D_{xz} , Figure S4 in the Supporting Information). While the HOMO-1 should thus be the nondegenerate 1S, the HOMO-1 is, in fact, approximately doubly degenerate due to mixing between the ligand orbitals and the valence d orbitals of the icosahedral gold core, and furthermore, the HOMO-1 (Figure S5 in the

Supporting Information) does not have super-atom 1S character, which was shown previously by Akola et al.⁶ Due to mixing in the sub-HOMO orbitals, a super-atom 1S orbital is difficult to identify but may be the HOMO-6 orbital. From the angular momentum selection rule $\Delta L = \pm 1$, the spectrum is expected to be dominated by transitions from the 1P HOMO to the 1D LUMO or LUMO+1 or to the 2S, which is the LUMO+2 in all the hybrid calculations and in the SAOP/BP86 calculation, and is the LUMO+3 in the

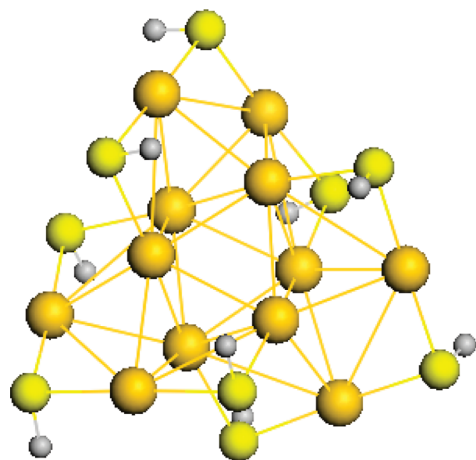


Figure 2. Structure for $\text{Au}_{12}(\text{SH})_9^{+1}$.

SAOP//X α calculation.⁸⁹ In addition, while the system does not have exact D_{2h} symmetry, C_i symmetry was enforced so only transitions that change parity are dipole allowed. The HOMO has a_u symmetry, and the LUMO, LUMO+1, and LUMO+2 (or LUMO+3 for SAOP// X α) are all a_g symmetry, so these transitions are all allowed. The HOMO-2, HOMO-5, and HOMO-7 (all triply degenerate) all have, like the HOMO, a_u symmetry so could also be involved in transitions to the low lying unoccupied orbitals, while HOMO-1, HOMO-3, HOMO-4, and HOMO-6 are a_g symmetry and thus are unlikely to be involved in the OPA spectrum. The SAOP, B3LYP, CAM, and mCAM were designed to correct for the asymptotic deficiencies of GGA functionals, such as BP86, but only the functionals that include exact exchange (B3LYP, CAM, and mCAM) yield significantly different HOMO–LUMO gaps (2.2–4.0 eV) than the BP86 gap of 1.30 eV, while the SAOP functional only increases the gap slightly compared to BP86. For these X-C functionals, the first peak in the spectrum (peak a) is due to HOMO \rightarrow LUMO transitions, but the linear response behavior of the hybrid functionals is significantly different from that of the nonhybrid functionals. In the SAOP calculations, the HOMO–LUMO gap of 1.48 eV is about 0.3 eV less than the corresponding experimental excitation energy, and linear response theory provides a small positive correction to produce a first peak at 1.63 eV, still about 0.2 eV below the experimental value. In the B3LYP calculation, the HOMO–LUMO gap is about 0.5 eV above the measured excitation energy, while in this linear response calculation, the negative contribution from the exact exchange component places the first peak in nearly exact agreement with experiment.

Peak b is dominated by HOMO \rightarrow LUMO+1 transitions in all the calculations. However, in the SAOP calculations, peak b also gets significant contributions from excited states dominated by HOMO-2 \rightarrow LUMO and HOMO \rightarrow LUMO+2 (or LUMO+3) transitions, while in the hybrid calculations, the excited states from these transitions are higher in energy and produce peak c. In the SAOP calculations, peak c is primarily from HOMO-5 \rightarrow LUMO transitions, while in the hybrid calculations, these transitions produce peak d.

2. TPA. The TPA for $\text{Au}_{25}(\text{SH})_{18}^{-1}$ was calculated using the BP86, B3LYP, and mCAM functionals with the LANL2DZ and SDD-DZ basis sets. The transition energies and the calculated TPA cross-sections are listed in Table S5 in the Supporting Information. As with the gold dimer systems, the high density of states makes resonance effects problematic, and a wide variation in peak TPA cross-sections is obtained. For each level of theory, the largest TPA cross-sections benefit from resonant enhancement from the first excited state, and the overall transition is



Where $n = 3-6$. In the super-atom theory, these final states are all 1F orbitals. As can be seen in the SOS expression, eq 2, TPA requires an intermediate state that has dipole-allowed transitions with both the initial and final states. Thus, with the excited states responsible for peak a in the OPA spectrum (which correspond to HOMO \rightarrow LUMO transitions) acting as the intermediate or “virtual” state in the TPA process, the TPA can be written like the analogous two-step OPA process:



thus obeying the $\Delta L = \pm 1$ rule for each step. As expected from this rule, transitions from the HOMO (1P) to the 1F orbitals were forbidden in OPA but are the strong TPA states. The B3LYP/SDD-DZ results include two TPA states with nearly the same transition energy as experiment, one of which (at 3.17 eV) has a TPA cross-section (620 000 GM) of a similar magnitude as experiment, while the other is significantly larger.

While the B3LYP/SDD-DZ calculation produced a number of measurable TPA transitions in the range of 3–4 eV, the calculated TPA to the states near 1.9 eV is very small. This is expected as these states are all A_u in the C_i symmetry of this molecule, and the selection rules require that TPA states be of A_g symmetry. While the experimental results include a TPA cross-section of 2700 GM at 1.92 eV, the experimental results are not TPA maxima but are measurements of TPA at the individual wavelengths of 1290 and 800 nm, corresponding to TPA transition energies of 1.92 and 3.10 eV. In the calculated TPA results reported here, a line width of 0.3 eV (fwhm) was used to allow for resolution of the different states, but the actual line width is unknown. Both experiment and our calculations indicate a large TPA peak of greater than 10^5 GM near a transition energy of 3.1 eV, and by using a line width in the range 0.5 to 1.0 eV, the tail of this large absorption band would extend to 1.9 eV, and the calculated TPA cross-section at this energy would be similar in magnitude to the reported experimental value. Thus, without some knowledge of the broadening factors for this system, only qualitative predictions of TPA can be calculated.

C. $\text{Au}_{12}(\text{SR})_9^{+1}$. The stability of thiolated gold clusters that obey the shell-filling super-atom theory, such as $\text{Au}_{25}(\text{SH})_{18}^{-1}$ for $n^* = 8$ and $\text{Au}_{102}(\text{SH})_{44}$ for $n^* = 58$, motivated a search for super-atom clusters with $n^* = 2$, and two structures were proposed.²⁴ While it was previously proposed⁹⁰ that the smallest thiolated gold clusters may be

Table 6. Calculated Au–Au Bondlengths in $\text{Au}_{12}(\text{SR})_9^{+1}$ (Å) as well as Average Deviation from the Previously Calculated Geometry²⁴

R =	H	H	H	H	CH ₃	CH ₃	CH ₃
	BP86/SR2	X α /ZORA	TPSS/ZORA	B3LYP/ZORA	X α /ZORA	TPSS/ZORA	TPSS ²⁴
Core							
1,2	2.882	2.837	2.776	2.881	2.867	2.872	2.875
2,3	2.882	2.837	2.776	2.881	2.866	2.872	2.875
1,3	2.882	2.837	2.776	2.881	2.867	2.874	2.875
4,5	2.882	2.837	2.776	2.881	2.868	2.865	2.875
5,6	2.882	2.837	2.776	2.881	2.869	2.864	2.875
4,6	2.882	2.837	2.776	2.881	2.866	2.861	2.875
1,4	2.821	2.758	2.944	3.027	2.745	2.793	2.813
1,6	2.748	2.715	2.921	2.852	2.709	2.745	2.760
2,4	2.748	2.715	2.921	2.852	2.710	2.746	2.760
2,5	2.821	2.758	2.944	3.027	2.744	2.795	2.813
3,5	2.748	2.715	2.921	2.852	2.709	2.745	2.760
3,6	2.821	2.758	2.944	3.027	2.745	2.792	2.813
av	2.833	2.787	2.854	2.910	2.797	2.819	2.831
av dev.	0.008	0.044	0.123	0.079	0.034	0.012	
Wings							
2,12	2.893	2.838	3.019	3.137	2.798	2.864	2.876
5,12	3.062	3.069	3.009	3.263	2.953	3.001	3.012
2,9	3.083	3.069	3.009	3.263	2.992	3.083	3.052
5,9	2.873	2.838	3.019	3.137	2.778	2.826	2.852
1,11	2.893	2.838	3.019	3.137	2.798	2.864	2.876
1,8	3.083	3.069	3.009	3.263	2.994	3.085	3.052
4,11	3.062	3.069	3.009	3.263	2.954	3.001	3.012
4,8	2.873	2.838	3.019	3.137	2.777	2.826	2.852
3,7	3.083	3.069	3.009	3.263	2.992	3.084	3.052
3,10	2.893	2.838	3.019	3.137	2.797	2.865	2.876
6,7	2.873	2.838	3.019	3.137	2.777	2.823	2.852
6,10	3.062	3.069	3.009	3.263	2.954	3.005	3.012
av	2.978	2.954	3.014	3.200	2.880	2.944	2.948
av dev.	0.030	0.031	0.089	0.252	0.068	0.020	

protected by $\text{SR}(\text{AuSR})_3$ (trimer) wings due to the need for this longer ligand to wrap around the smaller core, both structures use the dimer wings as in $\text{Au}_{25}(\text{SH})_{18}^{-1}$. The first proposed structure, $\text{Au}_8(\text{SR})_6$, has a tetrahedral Au_4 core and two $\text{SR}(\text{AuSR})_2$ wings. However, the optimized structure left the core exposed and was rejected as probably unstable.²⁴ The second proposed structure, $\text{Au}_{12}(\text{SR})_9^{+1}$, has an Au_6 octahedral core and three $\text{SR}(\text{AuSR})_2$ wings and has approximate C_3 symmetry (see Figure 2). Jiang et al.²⁴ optimized the geometry for this structure with R = Me using the TPSS⁹¹ X-C functional and evaluated the OPA using the PBE⁹² functional. The geometry for this system has been optimized in this study for R = H with BP86/SR2, X α /ZORA, and B3LYP/ZORA and for R = CH₃ with X α /ZORA and TPSS/ZORA, and the core bond lengths for each level of theory are given in Table 6. As with the Au_{25} cluster, shorter bondlengths are obtained in the gold core when the X α X-C functional is used in the optimization compared to the results with either the BP86 or the TPSS functionals. In the Au_{25} cluster, the X α geometry was in better agreement with the experimental geometry than the GGA geometries, and using the X α geometry in the TDDFT calculation produced a spectrum in better agreement with the experimental spectrum. While a compound with the formula $\text{Au}_{12}(\text{SR})_9$ has been identified in a study of polydispersed gold nanoclusters,⁹³ no experimental data on the structure is available for comparison. Thus, multiple geometries and functionals have been tested in the TDDFT calculations.

The excitation energy, oscillator strength, and primary orbital transition for each peak in the calculated OPA spectrum for $\text{Au}_{12}(\text{SR})_9^{+1}$ is listed in Table 7. The superatom theory can also be used for this system to a limited extent. With $n^* = 2$, the HOMO is the nondegenerate 1S, and the LUMO should be the triply degenerate 1P. However, ligand field splitting from the C_3 symmetry splits the 1P orbitals into a doubly degenerate LUMO and a nondegenerate LUMO+1. As expected, the first peak in the spectrum is due to HOMO \rightarrow LUMO transitions. Similar to the results from the Au_{25} cluster, when the hybrid functional B3LYP is used, the HOMO–LUMO gap is much larger (by about 1 eV) than when one of the nonhybrid GGA functionals is used, and the first peak in the linear response spectrum is blue-shifted by about 0.3 eV and has nearly twice the peak intensity. Both the blue-shift and intensity increase are even larger when mCAM is used.

For the higher states, the geometry used plays a major role. For the R = H system, when the BP86 geometry is used, the HOMO-1 \rightarrow HOMO gap is large (0.70, 0.97, and 1.08 eV for the PBE, B3LYP, and mCAM results, respectively), while the LUMO \rightarrow LUMO+1 gap is smaller (0.55, 0.59, and 0.63 eV). When the X α geometry is used, these gaps are reversed; the HOMO-1 \rightarrow HOMO gap is small (0.33, 0.59, and 0.77 eV), while the LUMO \rightarrow LUMO+1 gap is large (0.95, 1.01, and 1.04 eV). Thus, while the calculations with the BP86 geometry show a weak but distinct second peak at about 2.8 eV from the HOMO \rightarrow

Table 7. Calculated Absorption Peaks for $\text{Au}_{12}(\text{SR})_9^{+1}$ as well as the Dominant Orbital Transition and Corresponding Kohn–Sham gap (eV)^a

TDDFT	geometry						H-1 → H	L → L+1
R = Me								
PBE/tpz	TPSS	ΔE	1.811	2.536	3.016	3.651		
		ϵ_{max}	8.33×10^3	1.96×10^4	1.56×10^4	1.74×10^4		
			H → L	H-2 → L	H-5 → L			
PBE/sdd	TPSS	K–S gap	1.645	2.337			0.612	0.516
		ΔE	1.851	2.556	3.086	3.666		
		ϵ_{max}	9.36×10^3	2.03×10^4	1.85×10^4	1.71×10^4		
			H → L	H-2 → L	H-2 → L+1			
mCAM/sdd	X α	K–S gap	1.672	2.405	2.944		0.607	0.539
		ΔE	2.254	3.319	4.204			
		ϵ_{max}	2.03×10^4	4.21×10^4	3.56×10^4			
			H → L	H-2 → L	H-3 → L+1			
		K–S gap	3.415	4.536	5.416		1.050	0.811
R = H								
PBE/sdd	BP86	ΔE	1.975	2.790	3.330	4.015		
		ϵ_{max}	1.03×10^4	2.30×10^4	2.31×10^4	2.03×10^4		
			H → L	H-1 → L	H-3 → L+1	H-7 → L+1		
PBE/sdd	X α	K–S gap	1.778	2.480	3.164	3.937	0.702	0.546
		ΔE	1.916	2.711	2.981	3.606		
		ϵ_{max}	7.58×10^3	1.36×10^4	1.35×10^4	1.42×10^4		
			H → L	H-3 → L	H-4 → L	H-3 → L+1		
B3LYP/sdd	BP86	K–S gap	1.715	2.541	2.877	3.828	0.327	0.951
		ΔE	2.273	2.803	3.278	3.893		
		ϵ_{max}	1.75×10^4	4.54×10^3	3.82×10^4	3.55×10^4		
			H → L	H → L+1	H-2 → L	H-2 → L+1		
B3LYP/sdd	X α	K–S gap	2.797	3.391	3.883	4.477	0.966	0.594
		ΔE	2.225	3.250	4.175	4.390		
		ϵ_{max}	1.49×10^4	2.78×10^4	2.44×10^4	2.37×10^4		
			H → L	H-4 → L	H-4 → L+1	H → L+5		
mCAM/sdd	BP86	K–S gap	2.756	3.865	4.872		0.592	1.007
		ΔE	2.408	2.953	3.483	4.158		
		ϵ_{max}	2.16×10^4	6.03×10^3	4.33×10^4	3.70×10^4		
			H → L	H → L+1	H-2 → L			
mCAM/sdd	X α	K–S gap	3.583	4.213	4.774		1.078	0.630
		ΔE	2.377	2.877	3.447	3.852		
		ϵ_{max}	1.87×10^4	1.83×10^3	3.23×10^4	6.77×10^3		
			H → L	H-1 → L	H-4 → L	H-5 → L		
		K–S gap	3.555	4.322	4.901	5.129	0.767	1.039

^a Excitation energy in eV and extinction coefficient in liter/cm/mol.

LUMO+1 transition, when the X α geometry is used, the LUMO+1 orbital is too high in energy, and this peak is not found. In these calculations, HOMO-1 → LUMO transitions produce a weak shoulder on the next strong peak, which is dominated by transitions from HOMO-3 or HOMO-4 to the LUMO. Surprisingly, in both geometries, the LUMO+1 orbital is the $1P_z$ in the super-atom model, where the Z-axis is in the direction of the two unprotected faces in the octahedral core (perpendicular to the plane of the paper in Figure 2). In the more compressed X α geometry, the $1P_x$ and $1P_y$ LUMO orbitals are slightly stabilized due to more interaction with the ligands, while the $1P_z$ LUMO+1 is slightly destabilized, resulting in the larger LUMO → LUMO+1 gap. The X α geometry is also less spherically symmetric, which could also produce the greater deviation from the ideal super-atom model.

The use of a hybrid functional in the linear response calculation blue-shifts each of the major peaks relative to that obtained with a GGA functional. Using the X α geometry seems to slightly red-shift the lines and reduce the intensity relative to the BP86 geometry for both the B3LYP and mCAM functional. Using the methyl ligand instead of

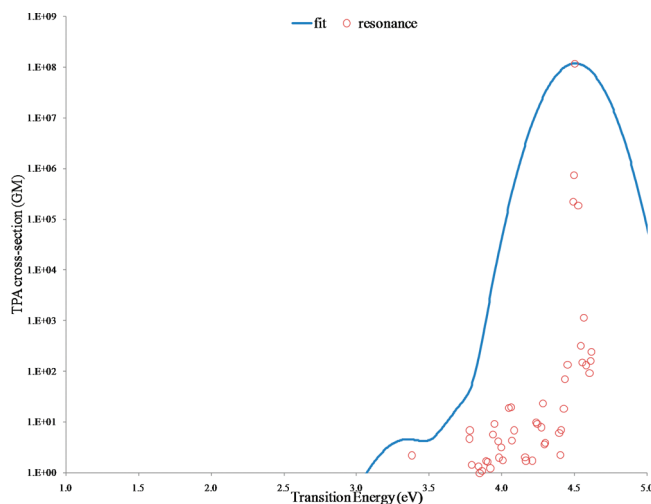


Figure 3. TPA for $\text{Au}_{12}(\text{SCH}_3)_9^{+1}$ using mCAM/SDD//X α .

hydrogen also causes a red-shift but increases the intensity. The density of states is quite high above 3 eV.

Figure 3 shows the calculated two-photon absorption using mCAM/SDD-DZ//X α . A Gaussian line shape was used with $\text{fwhm} = 0.3$ eV. The TPA seems to be dominated by

resonance effects from the first linear absorption line at 2.25 eV, resulting in a very large TPA at 4.5 eV. Due to the high density of states, several states have energies very close to resonance, including one at 4.501 eV with a TPA cross-section of 10^8 GM and three other states with TPA cross-sections greater than 10^5 GM. Compared to the TPA cross-sections calculated for the $\text{Au}_{25}(\text{SH})_{18}^{-1}$ cluster, the smaller $\text{Au}_{12}(\text{SR})_9^{+1}$ cluster has a smaller peak cross-section at a higher transition energy (for a given level of theory), which is consistent with the cluster-size scaling seen in the measured TPA cross-sections.^{2,4} No significant off-resonant TPA was found for $\text{Au}_{12}(\text{SR})_9^{+1}$.

V. Conclusions

The linear absorption spectrum has been calculated for the gold dimer and the monothiolated gold dimer anion using both time-dependent density functional theory (TDDFT) and the high-level completely renormalized equation-of-motion coupled cluster singles and doubles with perturbative triples (CR-EOMCCSD(T)) method. The CAMB3LYP and mCAM X-C functionals were found to give the best agreement with the high level of theory. The agreement between theory and experiment is not as good except when spin-orbit coupling effects are included in the calculations.

The calculated excitation energies of $\text{Au}_{25}(\text{SH})_{18}^{-}$ are in excellent agreement with experiment when the X α -optimized geometry is used with the B3LYP exchange-correlation (X-C) functional. Good agreement was also obtained with this geometry and the mCAM functional, and fair agreement was obtained when the BP86-optimized geometry was used with either the B3LYP, mCAM, or CAMB3LYP functional. The agreement was not as good when TDDFT was carried out with the BP86 functional, showing the importance of the asymptotic correction provided by the fraction of exact exchange in the hybrid functional. However, the systematic improvement in functionals that is required for achieving geometry optimization is still unclear.

The X-C functional used to optimize the geometry has a major effect on the calculated one-photon absorption (OPA) spectrum of $\text{Au}_{12}(\text{SR})_9^{+1}$. The more compressed and less-symmetric X α geometry increases the deviations from the super-atom model and significantly changes the calculated OPA spectrum, relative to the results using the BP86 X-C functional. This is in contrast to the $\text{Au}_{25}(\text{SR})_{18}^{-1}$ results, where use of this functional only compressed the gold core without a loss of spherical symmetry. The inclusion of exact exchange in the response calculation also has a significant effect on the $\text{Au}_{12}(\text{SR})_9^{+1}$ spectra.

For both $\text{Au}_{25}(\text{SH})_{18}^{-1}$ and $\text{Au}_{12}(\text{SR})_9^{+1}$, two-photon absorption (TPA) is dominated by resonance effects. The off-resonance TPA appears to be small. The large measured⁴ TPA for $\text{Au}_{25}(\text{SH})_{18}^{-1}$ of over 400 000 GM near a transition energy of 3.1 eV can be qualitatively explained by some of the calculated peaks, particularly when the B3LYP functional was used. However, the measured TPA of 2700 GM at a transition energy of 1.9 eV was not evident, although it could be the tail of the peak at 3.1 eV. Comparing the calculations on the clusters with 12 and 25 gold atoms, the smaller cluster has the smaller peak TPA cross-section at a higher transition

energy, as expected.⁴ This understanding could motivate ligand substitution with moieties that are known to exhibit large two-photon absorption.

Supporting Information Available: In addition to the data already mentioned, the geometries used in this study are given in Table S5. This information is available free of charge via the Internet at <http://pubs.acs.org>.

References

- (1) Jadzinsky, P. D.; Calero, G.; Ackerson, C. J.; Bushnell, D. A.; Kornberg, R. D. Structure of a Thiol Monolayer-Protected Gold Nanoparticle at 1.1 Å Resolution. *Science* **2007**, *318*, 430.
- (2) Ramakrishna, G.; Varnavski, O.; Kim, J.; Lee, D.; Goodson, T. Quantum-Sized Gold Clusters as Efficient Two-Photon Absorbers. *J. Am. Chem. Soc.* **2008**, *130*, 5032.
- (3) Zhu, M.; Aikens, C. M.; Hollander, F. J.; Schatz, G. C.; Jin, R. Correlating the Crystal Structure of A Thiol-Protected Au₂₅ Cluster and Optical Properties. *J. Am. Chem. Soc.* **2008**, *130*, 5883.
- (4) Ramakrishna, G.; Varnavskia, O.; Kimb, J.; Leeb, D.; Goodson, T.; Nonlinear Optical Properties of Quantum Sized Gold Clusters. In *Linear and nonlinear optics of organic materials VIII*, Proceedings of the SPIE The International Society for Optical Engineering, San Diego, CA, August 28, 2008; Jakubiak, R., Ed.; SPIE: Bellingham WA; Vol. 7049, p 70490L.1.
- (5) Si, S.; Gautier, C.; Boudon, J.; Taras, R.; Gladiali, S.; Burgi, T. Ligand Exchange on Au₂₅ Cluster with Chiral Thiols. *J. Phys. Chem. C* **2009**, *113*, 12966.
- (6) Akola, J.; Walter, M.; Whetten, R. L.; Hakkinen, H.; Gronbeck, H. On the Structure of Thiolate-Protected Au₂₅. *J. Am. Chem. Soc.* **2008**, *130*, 3756.
- (7) Heaven, M. W.; Dass, A.; White, P. S.; Holt, K. M.; Murray, R. W. Crystal Structure of the Gold Nanoparticle [N(C₈H₁₇)₄][Au₂₅(SCH₂CH₂Ph)₁₈]. *J. Am. Chem. Soc.* **2008**, *130*, 3754.
- (8) Walter, M.; Akola, J.; Lopez-Acevedo, O.; Jadzinsky, P. D.; Calero, G.; Ackerson, C. J.; Whetten, R. L.; Gronbeck, H.; Hakkinen, H. A unified view of ligand-protected gold clusters as superatom complexes. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 9157.
- (9) Knight, W. D.; Clemenger, K.; deHeer, W. A.; Saunders, W. A.; Chou, M. Y.; Cohen, M. L. Electronic Shell Structure and Abundance of Sodium Clusters. *Phys. Rev. Lett.* **1984**, *52*, 2141.
- (10) Ekardt, W. Dynamical Polarizability of Small Metal Particles: Self-Consistent Spherical Jellium Background Model. *Phys. Rev. Lett.* **1984**, *52*, 1925.
- (11) Ekardt, W. Work function of small metal particles: Self-consistent spherical jellium-background model. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1984**, *29*, 1558.
- (12) Jiang, D.-E.; Dai, S. From Superatomic Au₂₅(SR)₁₈- to Superatomic M@Au₂₄(SR)₁₈q Core-Shell Clusters. *Inorg. Chem.* **2008**, *48*, 2720.
- (13) Bishea, G. A.; Morse, M. D. Spectroscopic studies of jet-cooled AgAu and Au₂. *J. Chem. Phys.* **1991**, *95*, 5646.

- (14) Klotzbucher, W. E.; Ozin, G. A. Optical Spectra of Hafnium, Tungsten, Rhenium, and Ruthenium Atoms and Other Heavy Transition-Metal Atoms and Small Clusters ($Zr_{1,2}$, $Pd_{1,2}$, $Au_{1,2,3}$) in Noble Gas Matrices. *Inorg. Chem.* **1980**, *19*, 3767.
- (15) Simard, B.; Hackett, P. A. High Resolution Study of the (0, 0) and (1, 1) Bands of the $AO_u^+ - XO_g^+$ System of Au_2 . *J. Mol. Spectrosc.* **1990**, *142*, 310.
- (16) Ames, L. L.; Barrow, R. F. Rotational analysis of bands of the gaseous Au_2 molecule. *Trans. Faraday Soc.* **1967**, *63*, 39.
- (17) Morse, M. D. Clusters of Transition-Metal Atoms. *Chem. Rev.* **1986**, *86*, 1049.
- (18) James, A. M.; Kowalczyk, P.; Simard, B.; Pinegar, J. C.; Morse, M. D. The $A^1_u < X^0_g^+$ System of Gold Dimer. *J. Mol. Spectrosc.* **1994**, *168*, 248.
- (19) Harbich, W.; Fedrigo, S.; Buttet, J. Deposition of mass selected gold clusters in solid krypton. *J. Chem. Phys.* **1992**, *96*, 8104.
- (20) Ermler, W. C.; Lee, Y. S.; Pitzer, K. S. Ab initio effective core potentials including relativistic effects. IV. Potential energy curves for the ground and several excited states of Au_2 . *J. Chem. Phys.* **1979**, *70*, 293.
- (21) Das, K. K.; Balasubramanian, K. Spectroscopic Properties of Low-Lying Electronic States of Au_2 . *J. Mol. Spectrosc.* **1990**, *140*, 280.
- (22) Itkin, I.; Zaitsevskii, A. Quasirelativistic multipartitioning perturbation theory calculations on electronic transitions in Au_2 . *Chem. Phys. Lett.* **2003**, *374*, 143.
- (23) Wang, X.; Wan, X.; Zhou, H.; Takami, S.; Kubo, M.; Miyamoto, A. Electronic structures and spectroscopic properties of dimers Cu_2 , Ag_2 , and Au_2 calculated by density functional theory. *J. Mol. Struct. (Theochem)* **2002**, *579*, 221.
- (24) Jiang, D.-e.; Whetten, R. L.; Luo, W.; Dai, S. The smallest thiolated gold superatom complexes. *J. Phys. Chem. C* **2009**, *113*, 17291.
- (25) Gronbeck, H.; Walter, M.; Hakkinen, H. Theoretical Characterization of Cyclic Thiolated Gold Clusters. *J. Am. Chem. Soc.* **2006**, *128*, 10268.
- (26) Aikens, C. M. Effects of Core Distances, Solvent, Ligand, and Level of Theory on the TDDFT Optical Absorption Spectrum of the Thiolate-Protected Au_{25} Nanoparticle. *J. Phys. Chem. A* **2009**, *113*, 10811.
- (27) Negishi, Y.; Nobusada, K.; Tsukuda, T. Glutathione-Protected Gold Clusters Revisited: Bridging the Gap between Gold(I)-Thiolate Complexes and Thiolate-Protected Gold Nanocrystals. *J. Am. Chem. Soc.* **2005**, *127*, 5261.
- (28) Negishi, Y.; Chaki, N. K.; Shichibu, Y.; Whetten, R. L.; Tsukuda, T. Origin of Magic Stability of Thiolated Gold Clusters: A Case Study on $Au_{25}(SC_6H_{13})_{18}$. *J. Am. Chem. Soc.* **2007**, *129*, 11322.
- (29) Stener, M.; Nardelli, A.; Francesco, R. D.; Fronzoni, G. Optical Excitations of Gold Nanoparticles: A Quantum Chemical Scalar Relativistic Time Dependent Density Functional Study. *J. Phys. Chem. C* **2007**, *111*, 11862.
- (30) Rumi, M.; Ehrlich, J. E.; Heikal, A. A.; Perry, J. W.; Barlow, S.; Hu, Z.; McCord-Maughon, D.; Parker, T. C.; Rockel, H.; Thayumanavan, S.; Marder, S. R.; Beljonne, D.; Bredas, J.-L. Structure-property relationships for two-photon absorbing chromophores: bis-donor diphenylpolyene and bis(styryl)benzene derivatives. *J. Am. Chem. Soc.* **2000**, *122*, 9500.
- (31) Sutherland, R. L.; Nonlinear Absorption In *Handbook of Nonlinear Optics*; Thompson, B. J., Ed.; Marcel Dekker, Inc.: New York, 1996.
- (32) Day, P. N.; Nguyen, K. A.; Pachter, R. TDDFT Study of One- and Two-Photon Absorption Properties: Donor-pi-Acceptor Chromophores. *J. Phys. Chem. B* **2005**, *109*, 1803.
- (33) Day, P. N.; Nguyen, K. A.; Pachter, R. Calculation of two-photon absorption spectra of donor-acceptor compounds in solution using quadratic response time-dependent density functional theory. *J. Chem. Phys.* **2006**, *125*, 094103.
- (34) Day, P. N.; Nguyen, K. A.; Pachter, R. Calculation of One-Photon and Two-Photon Absorption Spectra of Porphyrins Using Time-Dependent Density Functional Theory. *J. Chem. Theory Comput.* **2008**, *4*, 1094-1106.
- (35) Orr, B. J.; Ward, J. F. perturbation theory of the non-linear optical polarization of an isolated system. *Mol. Phys.* **1971**, *20*, 513.
- (36) Birge, R. R.; Pierce, B. M. A theoretical analysis of the two-photon properties of linear polyenes and the visual chromophores. *J. Chem. Phys.* **1979**, *70*, 165.
- (37) Birge, R. R.; Bennett, J. A.; Hubbard, L. M.; Fang, H. L.; Pierce, B. M.; Kliger, D. S.; Leroi, G. E. Two-Photon Spectroscopy of all-trans-Retinal. Nature of the Low-Lying Singlet States. *J. Am. Chem. Soc.* **1982**, *104*, 2519.
- (38) Peticolas, W. L. Multiphoton spectroscopy. *Annu. Rev. Phys. Chem.* **1967**, *18*, 233.
- (39) McClain, W. M. Two-Photon Molecular Spectroscopy. *Acc. Chem. Res.* **1974**, *7*, 129.
- (40) Monson, P. R.; McClain, W. M. Polarization dependence of the two-photon absorption of tumbling molecules with application to liquid 1-chloronaphthalene and benzene. *J. Chem. Phys.* **1970**, *53*, 29.
- (41) Gold, A. *Proceedings of the International School of Physics*; Academic: New York, 1969.
- (42) McClain, W. M. Excited State Symmetry Assignment Through Polarized Two-Photon Absorption Studies of Fluids. *J. Chem. Phys.* **1971**, *55*, 2789.
- (43) Masthay, M. B.; Finsden, L. A.; Pierce, B. M.; Bocian, D. F.; Lindsey, J. S.; Birge, R. R. a theoretical investigation of the one- and two-photon properties of porphyrins. *J. Chem. Phys.* **1986**, *84*, 3901.
- (44) Albota, M.; Beljonne, D.; Bredas, J. L.; Ehrlich, J. E.; Fu, J. Y.; Heikal, A. A.; Hess, S. E.; Kogej, T.; Levin, M. D.; Marder, S. R.; McCord-Maughon, D.; Perry, J. W.; Rockel, H.; Rumi, M.; Subramaniam, G.; Webb, W. W.; Wu, X. L.; Xu, C. design of organic molecules with large two-photon absorption cross-sections. *Science* **1998**, *281*, 1653.
- (45) Karotki, A.; Drobizhev, M.; Dzenis, Y.; Taylor, P. N.; Anderson, H. L.; Rebane, A. Dramatic enhancement of intrinsic two-photon absorption in a conjugated porphyrin dimer. *Phys. Chem. Chem. Phys.* **2004**, *6*, 7.
- (46) Spangler, C. W.; Starkey, J. R.; Meng, F.; Gong, A.; Drobizhev, M.; Rebane, A.; Moss, B.; Targeted Two-photon Photodynamic Therapy for the Treatment of Subcutaneous Tumors. In *Optical Methods for Tumor Treatment and Detection: Mechanisms and Techniques in Photodynamic Therapy XIV*; Proceedings of the SPIE-The International Society for Optical Engineering, Bellingham, WA, April 5, 2005; Kessel, D., Ed.; SPIE: Bellingham, WA; Vol. 5689, p 141.

- (47) Kirkpatrick, S. M.; Baur, J. W.; Clark, C. M.; Denny, L. R.; Tomlin, D. W.; Reinhardt, B. R.; Kannan, R.; Stone, M. O. Holographic recording using two-photon-induced photopolymerization. *Appl. Phys. A: Mater. Sci. Process.* **1999**, *69*, 461.
- (48) Nguyen, K. A.; Day, P. N.; Pachter, R. Effects of solvation on one- and two-photon spectra of coumarin derivatives: A time-dependent density functional theory study. *J. Chem. Phys.* **2007**, *126*, 094303.
- (49) Wang, J.; Blau, W. J. Inorganic and hybrid nanostructures for optical limiting. *J. Opt. A: Pure Appl. Opt.* **2009**, *11*, 024001.
- (50) Castro, A.; Marques, M. A. L.; Romero, A. H.; Oliveira, M. J. T.; Rubio, A. The role of dimensionality on the quenching of spin-orbit effects in the optics of gold nanostructures. *J. Chem. Phys.* **2008**, *129*, 144110.
- (51) Douglas, M.; Kroll, N. M. Quantum Electrodynamical Corrections to the Fine Structure of Helium. *Ann. Phys.* **1974**, *82*, 89.
- (52) Hess, B. A. Relativistic electronic-structure calculations employing a two-component no-pair formalism with external-field projection operators. *Phys. Rev. A: At., Mol., Opt. Phys.* **1986**, *33*, 3742.
- (53) 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. General Atomic and Molecular Electronic Structure System. *J. Comput. Chem.* **1993**, *14*, 1347.
- (54) Agren, H., personal communication.
- (55) *Dalton, a molecular electronic structure program*, release 2.0; KTH: Stockholm, Sweden, 2005; <http://www.kjemi.uio.no/software/dalton/dalton.html>. Accessed March 4, 2005.
- (56) van Lenthe, E.; Baerends, E. J.; Snijders, J. G. Relativistic regular two-component Hamiltonians. *J. Chem. Phys.* **1993**, *99*, 4597.
- (57) van Lenthe, E.; Baerends, E. J.; Snijders, J. G. Relativistic total energy using regular approximations. *J. Chem. Phys.* **1994**, *101*, 9783.
- (58) *ADF*, release 2008.01; SCM, Theoretical Chemistry, Vrije Universiteit: Amsterdam, The Netherlands, 2008; <http://www.scm.com>. Accessed November 12, 2008.
- (59) teVelde, G.; Bickelhaupt, F. M.; Gisbergen, S. J. A. v.; Guerra, C. F.; Baerends, E. J.; Snijders, J. G.; Ziegler, T. Chemistry with ADF. *J. Comput. Chem.* **2001**, *22*, 931.
- (60) Guerra, C. F.; Snijders, J. G.; Velde, G. t.; Baerends, E. J. Towards an order-N DFT method. *Theor. Chem. Acc.* **1998**, *99*, 391.
- (61) Wang, F.; Ziegler, T. A simplified relativistic time-dependent density-functional theory formalism for the calculations of excitation energies including spin-orbit coupling effect. *J. Chem. Phys.* **2005**, *123*, 154102.
- (62) Wang, F.; Ziegler, T.; van Lenthe, E.; van Gisbergen, S.; Baerends, E. J. The calculation of excitation energies based on the relativistic two-component zeroth-order regular approximation and time-dependent density-functional with full use of symmetry. *J. Chem. Phys.* **2005**, *122*, 204103.
- (63) Becke, A. D. Density-functional exchange-energy approximation with correct asymptotic behavior. *Phys. Rev. A: At., Mol., Opt. Phys.* **1988**, *38*, 3098.
- (64) Perdew, J. P. Density-functional approximation for the correlation energy of the inhomogeneous electron gas. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1986**, *33*, 8822.
- (65) Lovallo, C. C.; Klobukowski, M. Improved Model Core Potentials for the Second- and Third-Row Transition Metals. *J. Comput. Chem.* **2004**, *25*, 1206–1213.
- (66) Slater, J. C. A simplification of the Hartree-Fock Method. *Phys. Rev.* **1951**, *81*, 385.
- (67) Tsuchiya, T.; Abe, M.; Nakajima, T.; Hirao, K. Accurate relativistic Gaussian basis sets for H through Lr determined by atomic SCF calculations with the third-order Douglas-Kroll approximation. *J. Chem. Phys.* **2001**, *115*, 4463.
- (68) Tsuchiya, T.; Abe, M.; Nakajima, T.; Hirao, K. Accurate relativistic Gaussian basis sets for H through Lr determined by atomic SCF calculations with the third-order Douglas-Kroll approximation; Riken: Wako, Saitama; http://www.riken.jp/qcl/publications/dk3bs/periodic_table.html. Accessed May 4, 2010.
- (69) Hay, P. J.; Wadt, W. R. Ab initio effective core potentials for molecular calculations. Potentials for K to Au including the outermost core orbitals. *J. Chem. Phys.* **1985**, *82*, 299.
- (70) Andrae, D.; Haeussermann, U.; Dolg, M.; Stoll, H.; Preuss, H. Energy-adjusted ab initio pseudopotentials for the second and third row transition elements. *Theor. Chim. Acta* **1990**, *77*, 123.
- (71) Bergner, A.; Dolg, M.; Kuchle, W.; Stoll, H.; Preuss, H. Ab-initio energy-adjusted pseudopotentials for elements of groups 13–17. *Mol. Phys.* **1993**, *80*, 1431.
- (72) Becke, A. D. B3LYP: Density-functional thermochemistry III. *J. Chem. Phys.* **1993**, *98*, 5648.
- (73) Lee, C.; Yang, W.; Parr, R. G. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron-density. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1988**, *37*, 785.
- (74) Vosko, S. H.; Wilk, L.; Nusair, M. Accurate spin-dependent electron liquid correlation energies for local spin density calculations: a critical analysis. *Can. J. Phys. Chem.* **1980**, *58*, 1200.
- (75) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. Ab initio Calculation of Vibrational absorption and circular dichroism spectra using density functional force fields. *J. Phys. Chem.* **1994**, *98*, 11623.
- (76) Yanai, T.; Tew, D. P.; Handy, N. C. A new hybrid exchange-correlation functional using the Coulomb-attenuating method (CAM-B3LYP). *Chem. Phys. Lett.* **2004**, *393*, 51.
- (77) Schipper, P. R. T.; Gritsenko, O. V.; van Gisbergen, S. J. A.; Baerends, E. J. Molecular calculations of excitation energies and (hyper)polarizabilities with a statistical average of orbital model exchange-correlation potentials. *J. Chem. Phys.* **2000**, *112*, 1344.
- (78) Piecuch, P.; Kucharski, S. A.; Kowalski, K.; Musial, M. Efficient computer implementation of the renormalized coupled-cluster methods: The R-CCSD[T], R-CCSD(T), CR-CCSD[T], and CR-CCSD(T) approaches. *Comput. Phys. Commun.* **2002**, *149*, 71.
- (79) Kowalski, K.; Piecuch, P. New coupled-cluster methods with singles, doubles, and noniterative triples for high accuracy calculations of excited electronic states. *J. Chem. Phys.* **2004**, *120*, 1715.

- (80) Wloch, M.; Gour, J. R.; Kowalski, K.; Piecuch, P. Extension of renormalized coupled-cluster methods including triple excitations to excited electronic states of open-shell molecules. *J. Chem. Phys.* **2005**, *122*, 214107.
- (81) Christiansen, O.; Koch, H.; Halkier, A.; Jorgensen, P.; Helgaker, T.; Meras, A. S. d. Large-scale calculations of excitation energies in coupled cluster theory: The singlet excited states of benzene. *J. Chem. Phys.* **1996**, *105*, 6921.
- (82) Olsen, J.; Jorgensen, P. Linear and nonlinear response functions for an exact state and for an MCSCF state. *J. Chem. Phys.* **1985**, *82*, 3235.
- (83) Hettema, H.; Jensen, H. J. A.; Jorgensen, P.; Olsen, J. Quadratic response functions for a multiconfigurational self-consistent field wave function. *J. Chem. Phys.* **1992**, *97*, 1174.
- (84) Furche, F. On the density matrix based approach to time-dependent density functional response theory. *J. Chem. Phys.* **2001**, *114*, 5982.
- (85) Salek, P.; Vahtras, O.; Guo, J.; Luo, Y.; Helgaker, T.; Agren, H. Calculations of two-photon absorption cross-sections by means of density-functional theory. *Chem. Phys. Lett.* **2003**, *374*, 446.
- (86) Agren, H.; Norman, P.; Baev, A. Multiphysics Modelling of Optical Materials. In *Optical Materials in Defence Systems Technology III*; Grote, J. G., Kajzar, F., Lindgren, M., Eds.; SPIE: Stockholm, Sweden, 2006; Vol. 6401, p 640103.
- (87) Zhao, Y.; Truhlar, D. G. The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: two new functionals and systematic testing of four M06-class functionals and 12 other functionals. *Theor. Chem. Acc.* **2008**, *120*, 215.
- (88) Grimme, S. Semiempirical GGA-Type Density Functional Constructed with a Long-Range Dispersion Correction. *J. Comput. Chem.* **2006**, *27*, 1787.
- (89) Aikens, C. M. Origin of Discrete Optical Absorption Spectra of M25(SH)18 - Nanoparticles (M = Au, Ag). *J. Phys. Chem. C* **2008**, *112*, 19797–19800.
- (90) Jiang, D.-e.; Chen, W.; Whetten, R. L.; Chen, Z. What Protects the Core When the Thiolated Au Cluster is Extremely Small. *J. Phys. Chem. C* **2009**, *113*, 16983–16987.
- (91) Tao, J. M.; Perdew, J. P.; Staroverov, V. N.; Scuseria, G. E. Climbing the Density Functional Ladder: Nonempirical Meta-Generalized Gradient Approximation Designed for Molecules and Solids. *Phys. Rev. Lett.* **2003**, *91*, 146401.
- (92) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* **1996**, *77*, 3865.
- (93) Zhang, Y.; Shuang, S.; Dong, C.; Lo, C. K.; Paa, M. C.; Choi, M. M. F. Application of HPLC and MALDI-TOF MS for Studying As-Synthesized Ligand-Protected Gold Nanoclusters Products. *Anal. Chem.* **2009**, *81*, 1676.
- (94) Aikens, C. M. Origin of Discrete Optical Absorption Spectra of M25(SH)18 - Nanoparticles (M = Au, Ag). *J. Phys. Chem. C* **2008**, *112*, 19797–19800.

CT100139T

Scaling Factors and Uncertainties for *ab Initio* Anharmonic Vibrational Frequencies

Russell D. Johnson III,^{*,†} Karl K. Irikura,[†] Raghu N. Kacker,[‡] and Rüdiger Kessel[‡]

Chemical and Biochemical Reference Data Division and Mathematical and Computational Sciences Division, National Institute of Standards and Technology, Gaithersburg, Maryland 20899-8320

Received May 10, 2010

Abstract: To predict the vibrational spectra of molecules, *ab initio* calculations are often used to compute harmonic frequencies, which are usually scaled by empirical factors as an approximate correction for errors in the force constants and for anharmonic effects. Anharmonic computations of fundamental frequencies are becoming increasingly popular. We report scaling factors, along with their associated uncertainties, for anharmonic (second-order perturbation theory) predictions from HF, MP2, and B3LYP calculations using the 6-31G(d) and 6-31+G(d,p) basis sets. Different scaling factors are appropriate for low- and high-frequency vibrations. The method of analysis is based upon the *Guide to the Expression of Uncertainty in Measurement*, published by the International Organization for Standardization (ISO). The data used are from the Computational Chemistry Comparison and Benchmark Database (CCCBDB), maintained by the National Institute of Standards and Technology, which includes more than 3939 independent vibrations for 358 molecules.

1. Introduction

One of the most popular uses of computational quantum chemistry models is to predict vibrational spectra. This is generally done in the harmonic approximation, in which the potential energy function (PEF) is taken as a truncated, second-order Taylor series. The neglect of higher-order curvature causes the predictions to deviate from experimental observations of fundamental frequencies. Moreover, the harmonic force constants are distorted by theoretical and numerical approximations inherent in the electronic structure calculations used to compute the PEF. As an approximate correction for these two sources of error, an empirical scaling factor (that is, a multiplicative correction) is usually applied to the theoretical frequencies. The value for the scaling factor is typically determined by least-squares fitting to a set of experimental vibrational frequencies.

It is possible to include higher-order terms in the Taylor series expansion of the PEF, although at significant compu-

tational cost. This is expected to improve the accuracy of the predictions. The anharmonic vibrational problem is usually solved using second-order perturbation theory with a harmonic reference wave function (VPT2), often with some treatment of Fermi resonances.¹ The other popular alternative is vibrational mean-field theory (VSCF) and its more sophisticated derivatives.^{2–4} Compared with VPT2, VSCF does not require a well-behaved Taylor series expansion for the PEF, is resistant to problems arising from near degeneracies, and is computationally expensive. Besides these popular approaches, specialized techniques are used for high-precision predictions for small molecules.^{5,6}

The present study, part of a series directed toward “virtual measurements”,⁷ is restricted to anharmonic vibrational frequencies predicted using VPT2. The goals are (1) to determine empirical scaling factors, along with their associated uncertainties, for some popular computational models and (2) to determine whether such scaling is helpful.

Our analysis is based upon the *Guide to the Expression of Uncertainty in Measurement* (GUM),⁸ which is a de facto international standard for quantifying the uncertainty in all

* Corresponding author e-mail: russell.johnson@nist.gov.

[†] Chemical and Biochemical Reference Data Division.

[‡] Mathematical and Computational Sciences Division.

Table 1. Repeatability of Replicate Vibrational Calculations for the 27 Vibrations of *n*-Propane, Using the 6-31G(d) Basis Sets

theory	HF	HF	B3LYP	B3LYP	B3LYP	B3LYP
convergence ^a	default	tight	tight	tight	tight	tight
grid ^b	n.a.	n.a.	(75, 302) ^c	(99, 590) ^d	(96, 32, 64)	(131, 974)
max. $\sigma(\omega)$ ^e	0.65 cm ⁻¹	0.02	0.02	0.02	0.03	0.02
max. $\sigma(\nu)$ ^f	11.7 cm ⁻¹	0.03	1.2	0.12	0.14	0.14
repetitions	998	998	998	465	128	176
$(\omega_i - \omega_{i,\text{ref}})$ ^g	(-0.1, 0.1)	0	(-4.4, 1.2)	0	(-0.1, 0.3)	(-0.1, 0.3)
$(\nu_i - \nu_{i,\text{ref}})$ ^h	(-3.6, 3.0)	0	(-6.5, 35.3)	0	(-4.5, 0.0)	(-1.7, 0.5)

^a Geometry convergence criteria (rms force: default = 3×10^{-4} au, tight = 1×10^{-5} au). ^b Integration grid for density functional calculations written either (radial, angular) or (radial, polar, azimuthal). ^c "Fine" pruned grid; software default. ^d "Ultrafine" pruned grid. ^e Largest standard deviation for a harmonic frequency, in cm⁻¹ units. ^f Largest standard deviation for an anharmonic fundamental frequency, in cm⁻¹ units. ^g Harmonic frequency shifts relative to reference calculation, written as a range (min, max) in cm⁻¹ units. Reference calculations are HF/tight for HF and B3LYP/tight/(99, 590) for B3LYP. ^h Anharmonic frequency shifts relative to reference calculation, as for harmonic frequencies.

kinds of measurements, including those determined from computational models.

2. Methodology

The procedure used here for determining scaling factors and their uncertainties closely follows that used for the scaling of harmonic vibrational frequencies,⁹ as slightly revised in a subsequent study of vibrational zero-point energies.⁷ Thus, it is presented here relatively briefly. Greater detail may be found in the earlier reports.

The scaling factor for a vibrational frequency of interest is obtained from a class of vibrational frequencies for which benchmark values are available. A class of vibrational frequencies is appropriate only if it meets three conditions.¹⁰ (1) The bias for the target frequency is believed to be of similar value to those in the class. (2) The (estimated) biases in the class have an approximately normal and acceptably narrow distribution. (3) The number of vibrational frequencies in the class is reasonably large. When classifying reference frequencies, technical knowledge and understanding are helpful. In the present study, benchmark, gas-phase, experimental values are taken from the Computational Chemistry Comparison and Benchmark Database (CCCB-DB).¹¹ The molecules included in this study are listed in the Supporting Information.

Suppose the value of the measurand (the target fundamental vibrational frequency of interest) is Y_0 and the corresponding value from a theoretical model is X_0 . Suppose the actual computed estimate of X_0 is x_0 with standard uncertainty $u(x_0)$. The *fractional bias* in x_0 is the ratio X_0/Y_0 . When the fractional bias in x_0 is believed to deviate significantly from unity, the GUM recommends that the estimate x_0 be scaled to counter its bias. The GUM requires that the measurement equation, which expresses the mathematical relationship between input quantities and the measurand, include an input variable for each component of the uncertainty. The measurement equation that corresponds to the fractional bias X_0/Y_0 is

$$Y_0 = C_0 X_0 \quad (1)$$

where C_0 is a scaling variable with a probability density function (pdf) representing the state of knowledge about the reciprocal (Y_0/X_0) of the fractional bias. Suppose c_0 and $u(c_0)$

are the expected value and the standard deviation of the state-of-knowledge pdf for C_0 . Then

$$y_0 = c_0 x_0 \quad (2)$$

The expected value $E(C_0) = c_0$ is a scaling factor applied to x_0 . By the common procedure of linear propagation of uncertainty,

$$u_r^2(y_0) \approx u_r^2(x_0) + u_r^2(c_0) \quad (3)$$

where we have taken the correlation coefficient $R(X_0, C_0) = r(x_0, c_0) = 0$ because the probability distributions for X_0 and C_0 are specified independently. The quantity $u_r(q) = u(q)/q$ is the relative standard uncertainty associated with the quantity q .

The uncertainty $u(x_0)$, which represents the dispersion in results that are obtained from nominally equivalent quantum chemistry calculations, arises from a variety of small contributions. For example, nonzero convergence thresholds create dependence upon the choice of the initial geometry and orbitals. We estimated the magnitude of this uncertainty using repeated calculations for propane (C₃H₈), which has 27 vibrational frequencies, starting from randomized initial coordinates (displacements up to $\pm 25\%$ of each z-matrix variable, from a uniform distribution). A calculation was discarded if its energy deviated by more than 10 standard deviations from the mean of all equivalent calculations. The results of this exercise are summarized in Table 1. The Hartree–Fock results indicate that tight geometry convergence criteria are necessary for these anharmonic calculations. The B3LYP results indicate that at least an "ultrafine" integration grid is also needed; values of $u(x_i)$ are only somewhat smaller, but the mean values are substantially different from those of the default grid. These conclusions affirm the original recommendations by Barone.¹²

The standard deviation of the distribution for the frequency x_i is an estimate of the standard uncertainty $u(x_i)$. The calculated anharmonic frequencies range approximately from 220 to 3100 cm⁻¹, so the calculations listed in Table 1 have values of $u_r(x_i)$ less than 0.0006, which can be neglected. This was found earlier for harmonic vibrational frequencies (using default criteria for geometry convergence).⁹ Taking $u_r(x_0) \approx 0$, from eq 3, we have $u(y_0) \approx y_0 u_r(c_0) = (y_0/c_0) u(c_0)$. Using eq 2 we have

$$u(y_0) \approx x_0 u(c_0) \quad (4)$$

Evaluated reference data are used to determine the scaling factor c_0 and the uncertainty $u(c_0)$, as described below. Then, the estimated value of a particular target vibrational frequency, y_0 , and its associated uncertainty, $u(y_0)$, can be determined from eqs 2 and 4, respectively.

Suppose the values for the fundamental vibrational frequencies in the specified class are Y_i and the corresponding values from a theoretical model are X_i , where $1 \leq i \leq m$, with m being the number of frequencies in the class. Estimates x_i (for X_i) are taken from actual calculations. Benchmark, experimental values z_i (estimates for Y_i) and their associated standard uncertainties $u(z_i)$ are taken from the CCCBDB.¹¹ The ratio $b_i = x_i/z_i$ is an estimate for the fractional bias X_i/Y_i , and the corresponding estimated scaling factor is

$$c_i = 1/b_i = z_i/x_i \quad (5)$$

Let C_i be a scaling variable for the fractional bias X_i/Y_i . The expected value and standard deviation of C_i are approximately c_i and $u(c_i)$, respectively. Taking $u(x_i) \approx 0$, as discussed above, we have

$$u(c_i) \approx u(z_i)/x_i \quad (6)$$

Define a nominal scaling variable $C_N = \sum_i k_i C_i$, where the k_i are normalized weights described below. Then the expected value of a state-of-knowledge pdf for C_N is

$$c_N = \sum_i k_i c_i \quad (7)$$

and the standard deviation is

$$u(c_N) = \left[\sum_i k_i^2 u^2(c_i) + \sum_i \sum_{j \neq i} k_i k_j u(c_i) u(c_j) r(c_i, c_j) \right]^{1/2} \quad (8)$$

When the state-of-knowledge pdf's of C_1, \dots, C_m are uncorrelated, the standard deviation (8) reduces to

$$u(c_N) = \left[\sum_i k_i^2 u^2(c_i) \right]^{1/2} \quad (9)$$

Since the bias in the unknown, target frequency is thought to be similar to the biases of the selected class of frequencies, it is reasonable to estimate the target scaling factor c_0 as $c_0 = c_N$. However, c_0 is likely to differ from c_N as suggested by the scatter of the estimated correction factors c_1, \dots, c_m about c_N . Therefore, the uncertainty $u(c_0)$ associated with $c_0 = c_N$ should be larger than the uncertainty $u(c_N)$. To determine $u(c_0)$, we need the following measurement equation for the correction variable C_0

$$C_0 = C_N + \delta C_0 \quad (10)$$

where δC_0 is a variable with a state-of-knowledge pdf for $C_0 - C_N$. A pdf for δC_0 is specified independently of the pdf for C_N ; therefore, they are uncorrelated. Suppose the expected value and standard deviation of δC_0 are δc_0 and $u(\delta c_0)$, respectively. Then

$$c_0 = c_N + \delta c_0 \quad (11)$$

and

$$u(c_0) = [u^2(c_N) + u^2(\delta c_0)]^{1/2} \quad (12)$$

The available information for specifying a state-of-knowledge pdf for δC_0 is the set of deviations of the estimated scaling factors c_i from the nominal scaling factor c_N . A useful class of pdf's that could be assigned to δC_0 are discrete probability distributions that assign a non-negative, normalized weight h_i to each deviation $(c_i - c_N)$.¹³ Then, the expected value is

$$\delta c_0 = \sum_i h_i (c_i - c_N) = \sum_i h_i c_i - c_N = \sum_i h_i c_i - \sum_i k_i c_i \quad (13)$$

and the variance is

$$u^2(\delta c_0) = \sum_i h_i [(c_i - c_N) - \delta c_0]^2 \quad (14)$$

We wish to use the nominal scaling factor c_N as the correction factor c_0 ; so from eq 11, we require that $\delta c_0 = 0$. From eq 13, this means that $\sum_i (h_i - k_i) c_i = 0$, which is satisfied if $h_i = k_i$ for all $i = 1, \dots, m$. Making these substitutions into eq 14 yields

$$u(\delta c_0) = \left[\sum_i k_i (c_i - c_N)^2 \right]^{1/2} \quad (15)$$

The usual practice when determining scaling factors is to fit the linear model

$$z_i = c_N x_i + e_i \quad (16)$$

by minimizing the least-squares objective function

$$\Delta^2 = \sum_i e_i^2 = \sum_i (z_i - c_N x_i)^2 \quad (17)$$

The well-known solution is

$$c_N = \frac{\sum_i x_i z_i}{\sum_i x_i^2} = \frac{\sum_i x_i^2 c_i}{\sum_i x_i^2} \quad (18)$$

which corresponds to eq 7 with weights

$$k_i = \frac{x_i^2}{\sum_i x_i^2} \quad (19)$$

To summarize, we have

$$c_0 = \frac{\sum_i x_i^2 c_i}{\sum_i x_i^2} \quad (20)$$

and, by combining eqs 9, 12, 15, and 19,

$$u(c_0) = \left[\frac{\sum_i x_i^4 u^2(c_i)}{(\sum_i x_i^2)^2} + \frac{\sum_i x_i^2 (c_i - c_0)^2}{\sum_i x_i^2} \right]^{1/2} \quad (21)$$

where c_i and $u(c_i)$ are given by eqs 5 and 6, respectively.

Electronic Structure Calculations. Anharmonic vibrational frequencies were computed using a partial quartic force field computed by using finite differences of analytical second derivatives.¹ All open-shell calculations were spin-unrestricted. Core orbitals were frozen, i.e., uncorrelated, in all MP2 calculations, including those involving group 1 (alkali) and group 2 (alkaline-earth) metals. Upon the basis of the results in Table 1, all calculations were done using tight geometry convergence criteria, and B3LYP calculations were done using the “ultrafine” grid (99 radial and 590 angular points). Because of software limitations, molecules were excluded if they possessed a 3-fold or higher symmetry axis. All computations were performed using the Gaussian 03 software package.^{14,15}

Where convenient, we use the symbol ω_{calc} to denote a computed, unscaled harmonic vibrational frequency and the symbol ν_{calc} to denote a computed, unscaled anharmonic frequency. The symbol ν_{expt} denotes an observed, experimental fundamental frequency.

3. Results and Discussion

Classification of Vibrational Frequencies. As stated in section 2, the scaling factor for a particular target frequency should be determined using reference data for an appropriate class of frequencies. When determining scaling factors for (harmonic) vibrational frequencies, it is traditional to consider all frequencies as a single class. Figure 1 shows a histogram of the estimated biases b_i , from eq 5, for anharmonic frequencies from the inexpensive HF/6-31G(d) model. Unlike the corresponding distribution for harmonic scaling factors,⁹ the histogram in Figure 1 is flat-topped. Thus, this class fails to satisfy the second criterion presented in section 2; that is, the distribution does not appear normal and reasonably narrow. We found it helpful to use two scaling factors, one for X–H stretches and one for all other vibrations. However, a simpler and nearly equivalent classification is between low and high (harmonic) frequencies. Earlier studies of harmonic frequencies have noted this distinction,^{16–22} often attributing it to anharmonicity. A plot of anharmonic bias against harmonic frequency, Figure 2, suggests that the boundary between “low” and “high” harmonic frequencies be set near 2700 cm^{-1} . Thus, we have chosen to use the value $\omega_{\text{calc}} = 2700 \text{ cm}^{-1}$ as the boundary for HF/6-31G(d). Histograms for both frequency ranges, displayed as line plots, are included in Figure 1. Analogous plots (see the Supporting Information) for the frequencies obtained from the other five models [B3LYP/6-31G(d), MP2/6-31G(d), HF/6-31+G(d,p), MP2/6-31+G(d,p), and B3LYP/6-31+G(d,p)] suggest setting the boundary between low and high frequencies at 2500 cm^{-1} for MP2 and 2600 cm^{-1} for B3LYP calculations. The MP2 and B3LYP models display smaller discrepancies between low and high frequencies than do the HF models. This suggests that the discrepancy in HF/6-31G(d) scaling

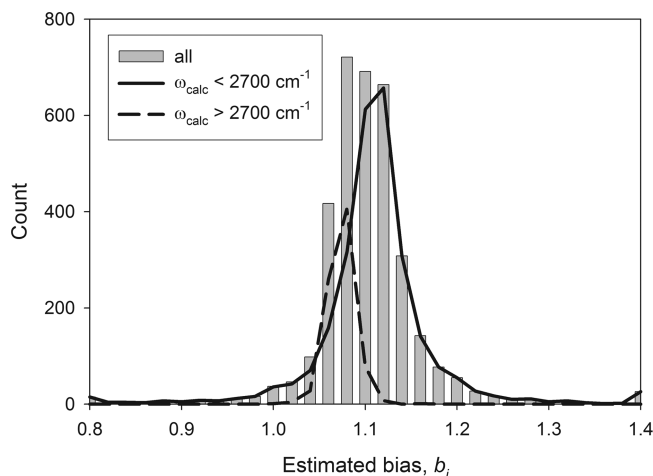


Figure 1. Histograms of estimated biases b_i for anharmonic HF/6-31G(d) calculations of 3446 frequencies of 215 molecules. The solid line is for low frequencies, the dashed line for high frequencies, and the bars for all frequencies combined.

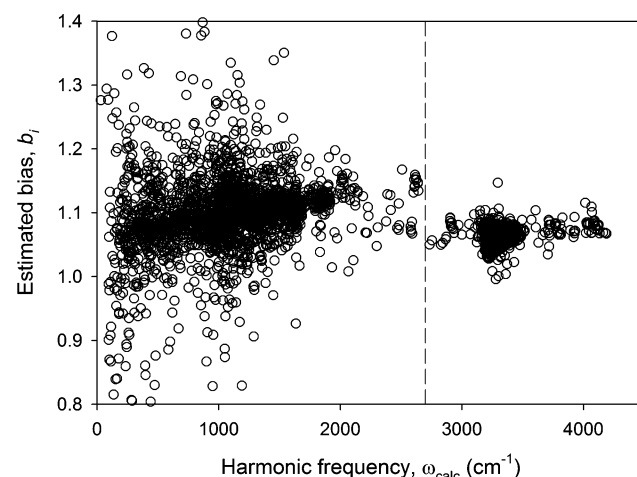


Figure 2. Distribution of estimated biases b_i for anharmonic HF/6-31G(d) calculations of 3446 frequencies of 215 molecules, plotted against calculated harmonic frequency, ω_{calc} . The dashed vertical line is drawn at 2700 cm^{-1} .

factors for high and low frequencies is primarily due to missing electron correlation. That is, X–H stretching vibrations are less affected by electron correlation than are other molecular vibrations, presumably because of their lower electron density.²³ As noted by a reviewer, the smaller values of $u(c_0)$ for high frequencies in all models (Table 2) can be explained by the greater similarity among the X–H stretching motions.

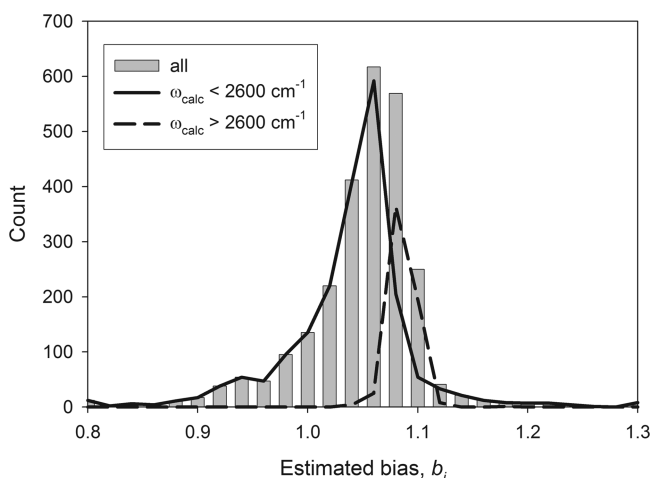
For MP2/6-31+G(d,p) calculations, the histogram of estimated bias for low-frequency vibrations is asymmetric (Figure 3). This indicates that more detailed classification is appropriate. However, we have not pursued this further in the present study.

Experimental Uncertainties Are Negligible. The first term of eq 21 is the contribution from the experimental uncertainties $u(z_i)$, via eq 6. As shown below, and as found previously for scaling harmonic frequencies,⁹ this term is negligible. We consider the popular B3LYP/6-31G(d) quantum chemistry model. Our data set for this model contains

Table 2. Anharmonic^a Scaling Factors, c_0 , and Their Associated Standard Uncertainties, $u(c_0)$

	$\omega_{\text{calc}} < \omega_{\text{high}}$	$\omega_{\text{calc}} > \omega_{\text{high}}$
HF/6-31G(d) ^b	0.9047 ± 0.0366 (2659)	0.9398 ± 0.0126 (783)
MP2/6-31G(d) ^c	0.9675 ± 0.0493 (2291)	0.9797 ± 0.0142 (673)
B3LYP/6-31G(d) ^d	0.9863 ± 0.0339 (2299)	1.0015 ± 0.0131 (675)
HF/6-31+G(d,p) ^e	0.9117 ± 0.0351 (2311)	0.9438 ± 0.0159 (675)
MP2/6-31+G(d,p) ^f	0.9822 ± 0.0352 (1998)	0.9698 ± 0.0117 (598)
B3LYP/6-31+G(d,p) ^g	0.9988 ± 0.0332 (2196)	1.0005 ± 0.0110 (647)

^a Using “tight” geometry optimization and “ultrafine” DFT integration grid. The number of vibrations in each data set is printed between parentheses. ^b $\omega_{\text{high}} = 2700 \text{ cm}^{-1}$. If all vibrations are considered a single class, $c_0 = 0.9285 \pm 0.0284$ (3442 frequencies). ^c $\omega_{\text{high}} = 2600 \text{ cm}^{-1}$. If all vibrations are considered a single class, $c_0 = 0.9759 \pm 0.0304$ (2964 frequencies). ^d $\omega_{\text{high}} = 2500 \text{ cm}^{-1}$. If all vibrations are considered a single class, $c_0 = 0.9967 \pm 0.0230$ (2974 frequencies). ^e $\omega_{\text{high}} = 2700 \text{ cm}^{-1}$. If all vibrations are considered a single class, $c_0 = 0.9335 \pm 0.0281$ (2986 frequencies). ^f $\omega_{\text{high}} = 2600 \text{ cm}^{-1}$. If all vibrations are considered a single class, $c_0 = 0.9735 \pm 0.0223$ (2596 frequencies). ^g $\omega_{\text{high}} = 2500 \text{ cm}^{-1}$. If all vibrations are considered a single class, $c_0 = 1.0000 \pm 0.0205$ (2843 frequencies).

**Figure 3.** Histograms of estimated biases b_i for anharmonic frozen-core MP2/6-31+G(d,p) calculations of 2599 frequencies of 176 molecules. The solid line is for low frequencies, the dashed line for high frequencies, and the bars for all frequencies combined.

$m = 495$ frequencies for which experimental uncertainties are known, primarily from the compilation by Shimanouchi.^{21,22} From eqs 20 and 21, we obtain $c_0 = 1.0001$ and $u(c_0) = 0.0203$ when all 495 vibrations are considered together. If the first term of eq 21 is ignored, we obtain the simpler, more approximate expression

$$u(c_0) \approx \left[\frac{\sum_i x_i^2 (c_i - c_N)^2}{\sum_i x_i^2} \right]^{1/2} \quad (22)$$

This yields a value of $u(c_0)$ that is smaller by only 0.000002. Similar results are obtained when only high or only low frequencies are considered. Thus, we conclude that eq 22 is adequate for estimating $u(c_0)$. Since uncertainties are not available for all of the experimental frequencies employed, eq 22 is used in all of the calculations described below.

Recommended Scaling Factors. Table 2 lists the estimated corrections for bias c_0 and estimated, associated

Table 3. Number of Anharmonic Frequency Predictions Improved (Worsened) Significantly^a by Scaling^b

	$\omega_{\text{calc}} < \omega_{\text{high}}$	$\omega_{\text{calc}} > \omega_{\text{high}}$	all ω_{calc} together
HF/6-31G(d)	2283 (218)	767 (9)	3129 (182)
MP2/6-31G(d)	1359 (492)	562 (81)	2002 (473)
B3LYP/6-31G(d)	889 (447)	0 (0)	55 (71)
HF/6-31+G(d,p)	1987 (198)	648 (16)	2687 (172)
MP2/6-31+G(d,p)	1060 (385)	556 (34)	1506 (598)
B3LYP/6-31+G(d,p)	0 (0)	0 (0)	0 (0)

^a By at least 10 cm^{-1} . ^b Computational details, ω_{high} values, and scaling factors from Table 2 and its footnotes.

standard uncertainties $u(c_0)$ for six theoretical models, i.e., combinations of theory and a one-electron basis set. In preparing the table, vibrations were excluded when the computed anharmonic correction exceeded 50%, that is, when $|\nu_{\text{calc}} - \omega_{\text{calc}}|/\omega_{\text{calc}} > 0.5$. Such large corrections are unusual and may indicate situations where vibrational perturbation theory is unreliable. If a single scaling factor is desired for the entire frequency range, it may be found among the footnotes to Table 2.

Is Scaling Helpful? Some of the scaling factors in Table 2 are within one standard uncertainty of unity, suggesting that scaling is not helpful. To examine this, we count the number of individual frequency predictions that are improved (or worsened) significantly by scaling, that is, the number of vibrations for which $|\nu_{\text{calc}} - \nu_{\text{expt}}|$ is less (or greater) than $|\nu_{\text{calc}} - \omega_{\text{calc}}|$ by at least 10 cm^{-1} . The results are listed in Table 3. Where the scaling factors differ noticeably from unity, many more frequencies are improved significantly by scaling than are degraded. We conclude that scaling is usually helpful. For B3LYP/6-31G(d) at higher frequencies ($\omega_{\text{calc}} > 2500 \text{ cm}^{-1}$) and for B3LYP/6-31+G(d,p) at all frequencies, the scaling factors are essentially unity, $c_0 = 1$. However, the values of $u(c_0)$ should not be approximated as zero. For example, high-frequency predictions using the B3LYP/6-31G(d) model may be scaled by the trivial factor of 1, with the resulting prediction having a relative standard uncertainty of 1.3%.

Effects of Basis Set. Although only two basis sets were studied, their comparison suggests that average scaling factors do not depend strongly upon the basis set, as was found earlier for harmonic scaling factors.⁹ The MP2 method is expected to be most sensitive because it uses virtual orbitals, whose number increases with basis-set size. The data of Table 2 substantiate this expectation. When the basis set is enlarged from 6-31G(d) to 6-31+G(d,p), low-frequency scaling factors increase by 0.007, 0.015, and 0.013 for HF, MP2, and B3LYP theories, respectively. The associated uncertainties change only for MP2. The corresponding high-frequency scaling factors change noticeably only for MP2, for which they decrease by 0.010.

Discrepancies between Software Versions. After this manuscript was completed, we learned that different versions of the software^{14,15} sometimes yield different results. To test whether this affects our conclusions, HF/6-31G(d) calculations were run on 215 molecules (3442 vibrational frequencies) using two software versions, B05 and E01. As before, frequencies were discarded when the anharmonic correction exceeded 50% of the harmonic frequency. Differences in ν_{calc}

between the different software versions ranged from -192 (for ν_2 of CH_2S) to $+247\text{ cm}^{-1}$ (for ν_8 of CH_2F_2), with a mean value of -0.8 cm^{-1} and standard deviation of 17.6 cm^{-1} . Values for 650 frequencies (19% of the set) differed by 10 cm^{-1} or more. However, the resulting scaling factors, 0.9047 ± 0.0366 and 0.9037 ± 0.0381 for low frequencies and 0.9398 ± 0.0126 and 0.9418 ± 0.0129 for high frequencies, do not differ significantly. Thus, our conclusions are unaffected.

Although tangential to the present study, we investigated this discrepancy somewhat further, as suggested by both reviewers. All harmonic frequencies are in excellent agreement between the two software versions. However, version E01 treats both vibrations (mentioned above) as affected by Fermi resonances, while version B05 does not. Moreover, there are some differences in cubic and quartic force constants. We were unable to devise any combination of options that would allow one version to reproduce the results of the other. As an independent test, the calculations on CH_2S and CH_2F_2 were repeated using the ACES2-MAB package.²⁴ This agreed with Gaussian03 for all harmonic frequencies and agreed with version B05 for ν_2 (CH_2S) and ν_8 (CH_2F_2). However, it disagreed seriously for ν_1 of both molecules [with B05, discrepancies of 193 cm^{-1} (CH_2S) and -247 cm^{-1} (CH_2F_2); with E01, discrepancies of 173 cm^{-1} (CH_2S) and -256 cm^{-1} (CH_2F_2)]. It is clear that the results are affected by choices made within different implementations. It is not clear whether any implementation is more “correct” than another.

4. Conclusions

Multiplicative scaling is an effective strategy for improving predictions from anharmonic vibrational calculations. The associated uncertainties reveal that the scaling factors have only two significant figures. However, we have reported four figures to conform to current, common practice for reporting harmonic scaling factors.

Slightly different scaling factors are recommended for low and high frequencies. The difference is greatest for HF calculations and smallest for B3LYP calculations. The uncertainties associated with low-frequency scaling factors are about 3 times as large as those for high-frequency scaling factors. For B3LYP/6-31+G(d,p) calculations, both high- and low-frequency scaling factors are negligibly different from unity.

Scaling factors depend only weakly upon the basis set. The scaling factors for high frequencies (X–H stretches) change less with the basis set than do the scaling factors for low frequencies.

Acknowledgment. Many of the calculations were performed on the “Biowulf” Linux cluster at the National Institutes of Health, Bethesda, MD (<http://biowulf.nih.gov>).

Supporting Information Available: Lists of molecules (one table) and histograms of estimated biases (four figures) and distributions of estimated biases (five figures) for theoretical models considered in this study. This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- (1) Barone, V. *J. Chem. Phys.* **2005**, *122*, 014108.
- (2) Matsunaga, N.; Chaban, G. M.; Gerber, R. B. *J. Chem. Phys.* **2002**, *117*, 3541–3547.
- (3) Christiansen, O. *Phys. Chem. Chem. Phys.* **2007**, *9*, 2942–2953.
- (4) Daněček, P.; Bouř, P. *J. Comput. Chem.* **2007**, *28*, 1617–1624.
- (5) Bowman, J. M.; Carter, S.; Huang, X. C. *Int. Rev. Phys. Chem.* **2003**, *22*, 533–549.
- (6) McCoy, A. B. *Int. Rev. Phys. Chem.* **2006**, *25*, 77–107.
- (7) Irikura, K. K.; Johnson, R. D., III; Kacker, R. N.; Kessel, R. *J. Chem. Phys.* **2009**, *130*, 114102.
- (8) *Guide to the Expression of Uncertainty in Measurement*, 2nd ed.; International Organization for Standardization (ISO): Geneva, Switzerland, 1995.
- (9) Irikura, K. K.; Johnson, R. D., III; Kacker, R. N. *J. Phys. Chem. A* **2005**, *109*, 8430–8437.
- (10) Irikura, K. K.; Johnson, R. D., III; Kacker, R. N. *Metrologia* **2004**, *41*, 369–375.
- (11) Johnson, R. D., III *NIST Computational Chemistry Comparison and Benchmark Database*, version 14; NIST Standard Reference Database Number 101, September 2006; National Institute of Standards and Technology: Gaithersburg, MD, 2006. <http://srdata.nist.gov/cccbdb/> (accessed March 31, 2009).
- (12) Barone, V. *J. Chem. Phys.* **2004**, *120*, 3059–3065.
- (13) Kacker, R. N.; Datta, R. U.; Parr, A. C. *J. Res. Natl. Inst. Stand. Technol.* **2003**, *108*, 439–446.
- (14) 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.; 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.: Pittsburgh, PA, 2003.
- (15) Certain commercial materials and equipment are identified in this paper in order to specify procedures completely. In no case does such identification imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the material or equipment identified is necessarily the best available for the purpose.
- (16) Scott, A. P.; Radom, L. *J. Phys. Chem.* **1996**, *100*, 16502–16513.
- (17) Bauschlicher, C. W.; Langhoff, S. R. *Spectrochim. Acta A* **1997**, *53*, 1225–1240.

- (18) Halls, M. D.; Velkovski, J.; Schlegel, H. B. *Theor. Chem. Acc.* **2001**, *105*, 413–421.
- (19) Yoshida, H.; Takeda, K.; Okamura, J.; Ehara, A.; Matsuura, H. *J. Phys. Chem. A* **2002**, *106*, 3580–3586.
- (20) Sinha, P.; Boesch, S. E.; Gu, C. M.; Wheeler, R. A.; Wilson, A. K. *J. Phys. Chem. A* **2004**, *108*, 9213–9217.
- (21) Shimanouchi, T. *Tables of Molecular Vibrational Frequencies, Consolidated Vol. I*; U.S. GPO: Washington, DC, 1972.
- (22) Shimanouchi, T. *J. Phys. Chem. Ref. Data* **1977**, *6*, 993–1102.
- (23) Irikura, K. K. *J. Phys. Chem. A* **1998**, *102*, 9031–9039.
- (24) Stanton, J. F.; Gauss, J.; Watts, J. D.; Szalay, P. G.; Bartlett, R. J.; Auer, A. A.; Bernholdt, D. E.; Christiansen, O.; Harding, M. E.; Heckert, M.; Heun, O.; Huber, C.; Johsson, D.; Jusélius, J.; Lauderdale, W. J.; Metzroth, T.; Michauk, C.; Price, D. R.; Ruud, K.; Schiffmann, F.; Tajti, A.; Varner, M. E.; Vázquez, J.; Almlöf, J.; Taylor, P. R.; Helgaker, T.; Jensen, H. J. A.; Jørgensen, P.; Olsen, J. *ACES II, vers. Mainz-Austin-Budapest*; Johannes Gutenberg-Universität: Mainz, Germany, 2005.

CT100244D

Sorting Out the Relative Contributions of Electrostatic Polarization, Dispersion, and Hydrogen Bonding to Solvatochromic Shifts on Vertical Electronic Excitation Energies

Aleksandr V. Marenich, Christopher J. Cramer,* and Donald G. Truhlar*

Department of Chemistry and Supercomputing Institute, University of Minnesota, 207 Pleasant Street S.E., Minneapolis, Minnesota 55455-0431

Received May 19, 2010

Abstract: Conventional polarized continuum model calculations of solvatochromic shifts on electronic excitation energies using popular quantum chemical programs (e.g., Gaussian or Turbomole) include the noninertial and inertial bulk-solvent polarization, which will be called electrostatics, but not dispersion interactions and specific effects like hydrogen bonding. For the $n \rightarrow \pi^*$ excitation of acetone in several solvents, we estimated the nonelectrostatic contributions in two ways: (i) the vertical excitation model (VEM) of Li et al. (*Int. J. Quantum Chem.* **2000**, *77*, 264), but updated to use TD-DFT corrected linear response with SMD atomic radii, and (ii) in the case of acetone in water, ensemble averaging over supermolecule calculations with up to 12 explicit solvent molecules selected from a molecular dynamics trajectory, with the explicit solvent surrounded by a continuum solvent. The TD-DFT VEM calculations carried out with the M06 density functional for 23 solvents result in a dispersion contribution to the red of 261–356 cm^{-1} and a hydrogen-bonding contribution to the blue of up to 289 cm^{-1} .

1. Introduction

It is well recognized that there are several contributions to solvatochromic effects. These include electric polarization of the solvent and electronic and geometric polarization of the solute, changes in dispersion, exchange repulsion, and cavitation, changes in first-solution-shell specific interactions (such as hydrogen bonding), and charge transfer between solute and solvent. In vertical excitation, it is assumed that vibrational and orientational polarization of the solvent and geometric relaxation of the solute and solvent do not have time to occur (and hence the solvent cavity size does not change if it is assumed to depend only on nuclear positions). Therefore, the following effects remain: (i) electronic polarization of the solvent, (ii) change in dispersion, (iii) change in exchange repulsion, and (iv) charge transfer; furthermore, there are contributions from (v) the interaction of the fixed slow polarization modes of the solvent with the changed electronic structure of the solute and (vi) changes

in the energy of hydrogen bonding, because the solvent configuration prior to electronic excitation is equilibrated with the electronic structure of the ground electronic state. We note that the major portion of the hydrogen bonding is electrostatic and is included in (i) and (v); when we refer to hydrogen bonding as a nonelectrostatic phenomenon, we mean the part that is not accounted for in (i) and (v). For example, this could be due to partial covalent character or due to the approximate nature of the treatment of dielectric polarization, including the model used for the solute–solvent boundary assumed in treating (i). In addition, because charge transfer from solute to solvent, that is, (iv), is usually not included explicitly, its especially large contribution to hydrogen bonding must also be considered as a specific first-solvation-shell effect.

Dielectric continuum models of the solvent can include (i) and (v). The primary goal of the present Article is to examine the importance of (ii) and (vi). We shall neglect (iii), which is a short-range effect that is hard to distinguish from the short-range component of (i) and (v). We also neglect explicit consideration of (iv), although it is partly

* Corresponding author e-mail: cramer@umn.edu (C.J.C.); truhlar@umn.edu (D.G.T.).

included in continuum dielectric treatments of (i) because of the empirical nature of the assumed solute–solvent boundary. For convenience, we will introduce the following short-hand names for the effects of interest: (i) fast polarization, (ii) dispersion, (v) slow polarization, and (vi) hydrogen bonding, but the reader should keep in mind that these are just labels for the wordier explanations given in the previous paragraph.

In the present Article, we study the solvatochromic shifts on the vertical $n \rightarrow \pi^*$ electronic excitation of acetone in several polar and nonpolar solvents by using two continuum solvation models for calculating the electrostatic component of the shifts as outlined below. Our choice of acetone is motivated by the abundance of experimental^{1–9} and theoretical data^{10–33} for its $n \rightarrow \pi^*$ transition in various solvents, which can be used for comparison to the gas phase. The dipole moment of acetone decreases upon the $n \rightarrow \pi^*$ electronic excitation; this, plus the slow response of hydrogen-bonding configurations to vertical excitation, leads to blue solvatochromic shifts in more polar solvents.²⁸ (A blue shift signifies an increase of the vertical excitation energy in solution relative to the gas phase and indicates that the solute is more favorably solvated in the ground state than in the excited state.) In nonpolar solvents, dispersion and shorter-range repulsion contributions to the corresponding vertical energy can dominate the bulk electrostatic contributions, thereby leading to red (bathochromic) shifts. Indeed, as noted by Rösch and Zerner¹² and earlier by Liptay,³⁴ the change in solute–solvent dispersion generally leads to red shifts upon vertical excitation. This is also discussed in a previous study from our group.¹⁹

In a previous study,¹⁹ we presented a treatment (based in part on earlier work³⁵) that included the fast and slow polarization, dispersion, and hydrogen bonding. It also included the coupling between the fast and slow polarization but otherwise neglected the coupling of the four identified effects with each other, which is an assumption but probably not a serious one. The previous model¹⁹ treated polarization effects with a two-time formulation of dielectric continuum theory³⁵ coupled to a configuration interaction model including single excitations (CIS) by intermediate-neglect-of-differential-overlap molecular orbital theory^{36,37} (in particular, INDO/S2³⁸ as incorporated in the ZINDO computer program³⁹). Since then, a similar two-time formulation has been coupled⁴⁰ to time-dependent density functional theory^{41–44} (TD-DFT), which provides a more generally accurate treatment of the electronic states of the solute, and hence potentially a more accurate treatment of the polarization. In the present Article, we re-examine the importance of dispersion and hydrogen bonding using TD-DFT to treat the polarization.

As an alternative to the treatment just described, in which the entire solvent is treated as a continuum, one can consider treating the most strongly coupled solvent molecules explicitly with the rest of the solvent treated as a continuum. We shall also consider this approach here, and we will compare it to the full continuum approach.

2. Theory

2.1. Electrostatics. The accurate theoretical description of solvatochromism, that is, condensed-phase effects on optical absorption and emission spectra, requires a proper account of nonequilibrium solute–solvent interactions, and, therefore, it poses an interesting challenge to theory.^{45–49} Whereas adiabatic transition energies correspond to both the initial and the final electronic states of the solute molecule being in equilibrium with a surrounding medium (solvent), vertical transition energies correspond to the final electronic state generated by photon absorption or emission being not equilibrated with the solvent environment. Within the continuum approximation for the polarizable solvent, the treatment of a solute molecule that undergoes an instantaneous change in its charge distribution via an electronic transition caused by photoabsorption or photoemission can be carried out using two time scales for solvent relaxation; these time scales correspond to the fast and slow components for the dynamic polarization response of the solvent.^{50,51} The fast polarization component is due to the response of the solvent electrons; this is sometimes called the noninertial polarization. The slow (inertial) response of the solvent is due to its nuclear motions. The fast (electronic) component of the solvent response is in instantaneous equilibrium with the nascent electronic state of the solute, whereas the inertial (nuclear) component is not. This leads to a nonequilibrium free energy that is similar to the nonequilibrium free energy introduced by Marcus in electron-transfer theory.⁵² In the present contribution, we consider dielectric medium effects on vertical electronic excitation from an equilibrated ground state to an excited electronic state.

Several dielectric continuum models using self-consistent reaction field (SCRF) theory have been adapted for calculation of vertical electronic transition energies based on two-time-scale response.^{19,35,48,49} There are also continuum models that treat multiple time scales using the complex frequency-dependent dielectric permittivity.^{53–56} In previous work,¹⁹ our group developed a two-response-time model of vertical excitation energies based on Aguilar et al.'s extension³⁵ of Marcus's theory of nonequilibrium free energies⁵² and on the distributed monopole representation of the solute charge density within the generalized Born approximation.^{57–62} A multiconfigurational self-consistent reaction field method that utilizes full multipole expansions of the solute charge field has also been developed.⁶³ Other methods for calculation of vertical electronic transition energies, for instance, those that use the continuous charge density (without approximating it by distributed point charges or multipoles) within the polarizable continuum model (PCM)^{64,65} framework have been reviewed in detail elsewhere.⁴⁸ The most recent nonequilibrium version^{66,67} of PCM utilizes the state-specific (SS) approach⁴⁸ in time-dependent density functional theory (TD-DFT) computations of both absorption and emission energies with analytical PCM/TD-DFT excited state energy gradients.⁶⁸ Another recent model exploits a less computationally expensive linear response (LR) approach, which has been corrected with an SS first-order perturbation correction⁶⁹ to approximate the SS solvent response within the

integral equation formalism polarizable continuum model (IEF-PCM).^{70–73} This method is called corrected linear response (cLR).⁶⁹ In addition, Chipman has derived a new formalism for the dielectric continuum treatment of vertical excitations within the surface and volume polarization for electrostatics (SVPE) framework.^{74,75} The SVPE method can be approximated through modification of the effective surface polarization without explicit volume polarization by the surface and simulation of volume polarization for electrostatics [SS(V)PE] method,^{76,77} which is essentially equivalent to IEF-PCM.⁷⁸

2.2. Specific Effects. The electrostatic treatments described above are labeled bulk electrostatics because the only solvent properties that they use are the bulk static dielectric constant and refractive index. (The square of the latter provides the dielectric constant at optical frequencies.) By invoking the continuum approximation of the solvent, one can eliminate the difficulties associated with the statistical sampling of solvent configuration space in the treatment of bulk electrostatic contributions to solvatochromic shifts. However, continuum models based on bulk electrostatics neglect solvent-structure effects and partial covalency associated with hydrogen bonding, and they do not treat solute–solvent dispersion forces or exchange repulsion of solute and solvent. (Note, however, that the solute–solvent dispersion interaction energy can be partially recovered in the LR approach by means of a term linear in the transition dipole moment between two electronic states of interest.^{66,79,80}) The assumption that the electrostatic interactions of the solute and the environment do not depend on the molecular structure of the solvent and that the dielectric response of the medium is isotropic outside the solute cavity can be inaccurate in cases when the solute molecule has strong specific interactions with one or a few solvent molecules in the first solvation shell. In addition, especially in nonpolar solvents, shorter-range (nonbulk) electrostatic effects and nonelectrostatic solute–solvent interactions can be equally or even more important than long-range electrostatic interactions.

In recent years, numerous attempts have been made to incorporate nonbulk electrostatic solvation effects into the treatment of excited electronic states within the continuum approximation. These can be based on atomic surface tensions representing an implicit solvent,⁴⁶ or they can include explicit consideration of one or more associated solvent molecules as part of the solute (which is then called a supersolute or supermolecule).^{28,46} Using this approach, one can account in part for changes in dispersion, exchange repulsion, or specific solute–solvent interactions (including hydrogen bonding and its associated solute–solvent charge transfer as well as other possible nonbulk-electrostatic and charge transfer effects) upon electronic excitation. However, in explicit models, one must explicitly average over a Boltzmann distribution of solvent orientations, and this rapidly becomes cumbersome as the number of explicit solvent molecules increases. An alternative approach involves addition of special nonelectrostatic corrections to the continuum description, for instance, an exchange repulsion term computed from classical pair potentials.²¹

One model we test here is the vertical electrostatic model (VEM) that was originally based on representing the solute by a set of distributed atomic monopoles within the generalized Born approximation and representing the solvent by its static and optical dielectric constants and augmenting the electrostatics by terms representing dispersion and hydrogen bonding.¹⁹ This model uses intermediate neglect of differential overlap for spectroscopy-parametrization 2 (INDO/S2) and configuration interaction wave functions constructed from single excitations (CIS).^{36–38} This model was implemented using the SM5.42 implicit solvation model,⁸¹ and the full vertical excitation model was abbreviated VEM42/INDO/S2 or VEM42. The VEM42 ground-state wave function is evaluated by a closed-shell semiempirical Hartree–Fock self-consistent reaction field (SCRf) calculation with an electrostatic reaction field corresponding to equilibrium solvation. The excited-state wave function is obtained within an iterative configuration interaction calculation including all single excitations with a two-response-time electrostatic reaction field corresponding to equilibrated fast (electronic) solvent response and nonequilibrated Franck–Condon-defined slow (nuclear) solvent response.¹⁹

Another continuum electrostatic model applied in the present study is the IEF-PCM model based on time-dependent density functional theory (TD-DFT) within the corrected linear response approach of Caricato et al.⁶⁹ to treat nonequilibrium solvation effects on vertical excitation spectra. This method is further abbreviated cLR/PCM/TD-DFT or cLR. Here, this kind of calculation is incorporated in the VEM as an improved bulk-electrostatic component that is augmented with contributions accounting for solute–solvent dispersion and hydrogen-bonding effects in the same fashion as in the original VEM model.

We have also studied the effect of making up to 12 water molecules explicit to obtain a better understanding of the way that the first solvation shell affects the magnitude of the solvatochromic shift in aqueous solution. Relevant supermolecular structures were generated from a molecular dynamics trajectory.

In the cLR calculations, we have systematically tested the dependence of calculated solvatochromic shifts on the values of intrinsic atomic Coulomb radii used for construction of the boundary between the solute cavity and the solvent continuum in the bulk electrostatic part of the calculations. In addition to the nine solvents tested in our previous work,¹⁹ we have used the experimental reference data on the $n \rightarrow \pi^*$ absorption of acetone in 23 solvents⁹ derived by Renge with the use of a new “band-halving” method for reliable determination of solvent shifts for poorly defined maxima of broad spectral envelopes (Renge corrected inconsistencies in the solvatochromic shifts of acetone existing in older literature for less polar media due to inconsistencies in locating the maxima of broad spectra).⁹

2.3. Detailed Theory. The polarization response of a medium described by the frequency-dependent permittivity $\epsilon(\omega)$ can be decomposed, within either the Pekar⁸² or the Marcus⁴⁸ partition, into two terms, fast and slow, as discussed above. The fast polarization component, which is physically due to electronic relaxation, can be described in terms of

the optical dielectric constant ϵ_{opt} equal to the square of the solvent refractive index, n^2 , at an optical frequency (ω_{opt}) at which the slow (mostly orientational) polarization can no longer follow the changes in the field.^{45–48} The slow polarization component, which physically requires nuclear motion of the solvent, can be identified by subtracting the fast component from the total polarization, which depends on the solvent static dielectric constant ϵ . The two partitions differ in that in the Pekar decomposition, the part of the fast polarization that is in equilibrium with nuclear polarization is included in the slow response (along with the polarization due to nuclear movement), but in the Marcus decomposition it is considered as part of the fast response.^{48,51} In the limit of linear response, which is assumed, where appropriate, in both the VEM42 and the cLR models, the two decompositions, each used with its corresponding expression for the nonequilibrium free energy, yield identical reaction fields and identical solvatochromic shifts.⁵¹

If the solvent is represented as a dielectric continuum, the electronic free energy G of the solute molecule in the ground state (G) can be expressed as

$$G_{\text{eq}}^{\text{G}} = \langle \Psi^{\text{G}} | \hat{H}_0 | \Psi^{\text{G}} \rangle + \frac{1}{2} \langle \Psi^{\text{G}} | V_{\text{fast}}^{\text{G}} | \Psi^{\text{G}} \rangle + \frac{1}{2} \langle \Psi^{\text{G}} | V_{\text{slow}}^{\text{G}} | \Psi^{\text{G}} \rangle \quad (1)$$

where \hat{H}_0 is the solute electronic Hamiltonian in the gas phase, Ψ^{G} is the solute electronic wave function in solution (because both the solvent and the solute are mutually polarized, the solute electronic wave function in solution differs from that in the gas phase), and V is the reaction potential induced by the polarization of the dielectric medium. The reaction field is expressed as a sum of its fast and slow components. The subscript “eq” in eq 1 indicates that both the fast and the slow components of the solvent reaction field are equilibrated with the solute charge density for the given electronic state. This equilibration can be achieved routinely through a self-consistent reaction field procedure.⁸³ Note that the VEM42 model based on the generalized Born approximation constructs the reaction potential V in terms of distributed monopoles (in particular, partial atomic charges), whereas the cLR/IEF-PCM model (or PCM in general) constructs the reaction field in terms of apparent surface charges distributed over the surface of the cavity containing the solute.

For a vertically excited solute molecule in a continuum solvent, the equilibrium free energy of eq 1 is replaced by a more complicated two-response-time nonequilibrium free energy. The detailed equations are given elsewhere^{19,35,48,51} for the Marcus decomposition; the more compact expression obtained with the Pekar decomposition is⁴⁸

$$G_{\text{neq}}^{\text{E}} = \langle \Psi^{\text{E}} | \hat{H}_0 | \Psi^{\text{E}} \rangle + \frac{1}{2} \langle \Psi^{\text{E}} | V_{\text{fast}}^{\text{E}} | \Psi^{\text{E}} \rangle + \langle \Psi^{\text{E}} | V_{\text{slow}}^{\text{G}} | \Psi^{\text{E}} \rangle - \frac{1}{2} \langle \Psi^{\text{G}} | V_{\text{slow}}^{\text{G}} | \Psi^{\text{G}} \rangle \quad (2)$$

The subscript “neq” indicates a nonequilibrium solvation regime in which only the $V_{\text{fast}}^{\text{E}}$ component is in equilibrium with the solute charge density, whereas the $V_{\text{slow}}^{\text{G}}$ component in eq 2 depends on the charge density in the ground state.

The calculation of $G_{\text{neq}}^{\text{E}}$ can be carried out using a two-step procedure: (i) an equilibrium calculation on the ground state (eq 1) to obtain $V_{\text{slow}}^{\text{G}}$ either in the form of partial atomic charges (SM5.42^{84–86} or other^{87–89} generalized Born models) or in the form of apparent surface charges (PCM); and (ii) a nonequilibrium calculation on the excited state (eq 2) using the fixed $V_{\text{slow}}^{\text{G}}$ component while relaxing Ψ^{E} and $V_{\text{fast}}^{\text{E}}$.

The potential $V_{\text{fast}}^{\text{E}}$ in eq 2 depends on the excited-state wave function of the solute in solution (or the solute’s density matrix in solution), and equilibrium between the charge density of the solute excited state and the fast degrees of freedom of the solvent can be obtained through a self-consistent (iterative) procedure using a state-specific approach such as the CIS model in VEM42. These self-consistency iterations are performed for the INDO/S2/CIS model, but for TD-DFT we employ the cLR method, which uses a first-order perturbation correction to obtain an approximation to state-specific solvation in which the slow response of the solvent is in equilibrium with the ground state and the fast response of the solvent is in equilibrium with the specific excited state of interest. This is particularly suitable for use with TD-DFT in which excitation energies are approximated as the poles of frequency-dependent linear response functions of the solute in the ground state, thereby avoiding calculation of the relaxed density matrix of an excited state.

The VEM42 model uses CM2 class IV charges^{81,90} obtained from INDO/S2 solute wave functions and the SM5.42 intrinsic Coulomb radii^{84,85} to construct molecular cavities in electrostatic calculations. The solvent is represented by its static and dynamic (optical) dielectric constants. All of the VEM42 calculations were carried out using a locally modified version⁹¹ of the ZINDO computer program.³⁹

All of the cLR/PCM calculations were performed using time-dependent density functional theory^{41–44} (TD-DFT) with the MG3S⁹² basis set. The Gaussian 09 program⁹³ (revision A.02) was employed. The cLR method is also called “StateSpecific”, “StateSpecificPerturbation”, or “SSPerturbation” in the online manual⁹⁴ of Gaussian 09. We employed the Gaussian 09 default settings for nonequilibrium solvation calculations, including the default tessellation algorithm and settings for the cavity definition, but the default intrinsic Coulomb radii were overridden by using the modify sph option to specifically assign values of radii and scaling factors. The tested solvents were defined using additional input keywords, eps and epsinf for static dielectric constant (ϵ) and for optical dielectric constants ($\epsilon_{\text{opt}} = n^2$), respectively, where n is the solvent refractive index.

Unless noted otherwise, all of the vertical excitation energies of acetone in solution were computed at the molecular geometry of acetone in its ground electronic state optimized in a given solvent with the use of the equilibrium SMD solvation model⁹⁵ implemented in Gaussian 09 using the M06-2X^{96,97} density functional with the MG3S⁹² basis. The corresponding gas-phase vertical excitation energy was computed using the M06-2X/MG3S geometry optimized in the gas phase.

Note that the LR method may include some solvent–solute dispersion,^{66,79,80} but an explicit treatment, not used here,

that includes dispersion fully is more complicated.⁸⁰ However, in practice the cLR algorithm⁶⁹ yields results in reasonable agreement with a method⁶⁹ that does not contain any solvent–solute dispersion, and so we proceed under the assumption that the electrostatic calculations do not contain dispersion nor does the bulk electrostatics portion contain specific-solvent effects like hydrogen bonding. The VEM42 and cLR/PCM bulk electrostatic contributions to the solvatochromic shifts were therefore augmented with one-parameter empirical corrections accounting respectively for solute–solvent dispersion and hydrogen-bonding effects, which are the most significant nonelectrostatic contributions. According to ref 19, the total solvatochromic shift is defined as follows:

$$\Delta\omega = \Delta\omega_{\text{EP}} + \Delta\omega_{\text{D}} + \Delta\omega_{\text{H}} \quad (3)$$

where $\Delta\omega_{\text{EP}}$ is the electrostatic contribution to the corresponding solvatochromic shift (EP denotes electronic and polarization, i.e., solute electronic energy plus net electric polarization free energy where “net” refers to the accounting, where appropriate, in the derivation of eq 2 for the work done in solvent polarization); $\Delta\omega_{\text{D}}$ is the shift due to solvent–solute dispersion; and $\Delta\omega_{\text{H}}$ is the shift due to hydrogen bonding. The electrostatic contribution is evaluated as follows:

$$\Delta\omega_{\text{EP}} = \omega_{\text{EP}}(\text{gas}) - \omega_{\text{EP}}(\text{liq}) \quad (4)$$

where $\omega_{\text{EP}}(\text{gas})$ and $\omega_{\text{EP}}(\text{liq})$ are the electrostatic contributions to the vertical excitation energies, respectively, in the gas phase and in solution. According to the established sign convention⁹⁸ that we use here, a red (bathochromic) shift is called positive (it corresponds to a decrease in frequency and increase in wavelength), and a blue (hypsochromic) shift is called negative (with an increase in frequency and decrease in wavelength). For solvent–solute dispersion ($\Delta\omega_{\text{D}}$), we assume a characteristic contribution for a given transition that depends on only the solvent as follows:^{1,19}

$$\Delta\omega_{\text{D}} = D \frac{n^2 - 1}{2n^2 + 1} \quad (5)$$

where D is a characteristic constant for this particular transition of acetone, and n is the refractive index of the solvent.¹⁹

The hydrogen-bonding contribution to the $n \rightarrow \pi^*$ excitation of acetone is likely to be dominated by proton donation from the solvent to the carbonyl oxygen, and this effect should correlate with the proton-donor capability of the solvent.¹⁹ Therefore, we assume

$$\Delta\omega_{\text{H}} = H\alpha \quad (6)$$

where H is a model parameter, and α is Abraham’s hydrogen-bond acidity parameter^{99–102} for a given solvent.

The solvent–solute dispersion contribution (eq 5) and the hydrogen-bonding contribution (eq 6) are added post-SCF, and they do not affect the solute charge distribution.

The model parameters D in eq 5 and H in eq 6 are optimized simultaneously by a least-squares fitting of

Table 1. Vertical Excitation Energies (ω , cm^{-1}) for the $n \rightarrow \pi^*$ Transition of Acetone in the Gas Phase^a

method	ω
CIS/MG3S	42 768
TD-B3LYP/MG3S	36 147
TD-M06/MG3S	36 067
TD-M06-HF/MG3S	28 144
TD-M06-L/MG3S	38 685
TD-M06-2X/MG3S	34 324
INDO/S2	33 055
INDO/S2	33 237 ^b
experiments	36 232, ^c 36 100, ^d 35 975 ^e
experimental average	36 102

^a Vertical excitation energies were calculated in the present work at the M06-2X/MG3S geometry optimized in the gas phase unless noted otherwise. ^b Reference 19. ^c Reference 8. ^d Reference 6. ^e Reference 9.

theoretical $\Delta\omega$ values to the corresponding experimental ones⁹ over the set of 23 solvents.

We use the values of solvent dielectric constant ϵ , refractive index n , and Abraham’s hydrogen-bond acidity parameter α presented in the Minnesota Solvent Descriptor Database¹⁰³ for all solvents, except for perfluoro-*n*-octane, tetraethoxysilane, *tert*-butyl chloride, and propylene carbonate. For these four solvents, which are not in the database, we use the values of ϵ and n from ref 9, and we assume that α is zero for these solvents.

3. Results and Discussion

3.1. Treatment with No Explicit Solvent Molecules.

First, we performed calculations of the vertical $n \rightarrow \pi^*$ excitation energies of acetone in the gas phase using TD-DFT with various density functionals (B3LYP,^{104–107} M06,^{96,97} M06-HF,^{97,108} M06-L,^{97,109} and M06-2X^{96,97}) and the MG3S basis set⁹² and using the CIS method^{36,110} both with the MG3S basis set⁹² and with the semiempirical INDO/S2 model.^{19,38} The computational results are compared to the available experimental data in Table 1. The ab initio CIS calculations overestimate the gas-phase excitation energy by $\sim 6670 \text{ cm}^{-1}$ on average, whereas the TD-M06-HF method underestimates it by $\sim 7960 \text{ cm}^{-1}$. The B3LYP and M06 density functionals provide the most accurate predictions of the vertical $n \rightarrow \pi^*$ excitation energies of acetone in the gas phase with respect to the corresponding experimental energies: compare 36 147 (B3LYP), 36 067 (M06), and 35 975–36 232 cm^{-1} (experiment). For this reason, we selected B3LYP and M06 for use in the cLR/PCM/TD-DFT calculations on the vertical $n \rightarrow \pi^*$ excitations of acetone in solution.

Table 2 shows six types of intrinsic atomic Coulomb radii used in the bulk-electrostatic calculations: Bondi’s van der Waals radii,¹¹¹ radii optimized for the SM5.42 solvation model,^{84,85} radii optimized for the SMD solvation model,⁹⁵ united atom topological models UA0¹¹² and UAHF,¹¹³ and universal force field (UFF) radii.¹¹⁴ The SM5.42 radii were tested using both the VEM42 and the cLR/PCM methods, whereas the other types were tested only with cLR/PCM. The UFF radii were also tested with the scaling factor of 1.1 and 1.3. Note that the UFF radii scaled by 1.1 are the default radii for excited-state solvation energy calculations in Gaussian 09.

Table 2. Tested Sets of Intrinsic Coulomb Radii (Å)

radii	hydrogen	carbon	oxygen
Bondi ^a	1.2	1.7	1.52
SM5.42 ^b	0.91	1.78	1.6
SMD ^c	1.2	1.85	1.52–2.294
UA0 ^{d,e}	n/a	1.925, ^f 2.525 ^g	1.75
UAHF ^{e,h}	n/a	1.68, ^f 1.95 ^g	1.59
UFF ^{e,i}	1.443	1.926	1.75
1.1×UFF ^j	1.5873	2.1186	1.925
1.3×UFF ^j	1.8759	2.5038	2.275

^a Bondi's values of van der Waals radii.¹¹¹ ^b Reference 84. ^c Reference 95. The SMD radius for O is defined as a function of Abraham's hydrogen-bond acidity parameter (α) for a given solvent. It is equal to 1.52 Å for any solvent with $\alpha \geq 0.43$, and it is equal to 2.294 Å for any solvent with $\alpha = 0$. For the solvents with $0 < \alpha < 0.43$, we used the following oxygen radii (in Å): 2.114 (dichloromethane and 1,2-dichloroethane), 2.096 (*cis*-dichloroethylene), 2.222 (acetone), and 2.168 (acetonitrile). ^d The united atom topological model UA0 of the acetone molecule with hydrogen atoms summed into methyl carbon atoms.¹¹² ^e Values given according to Gaussian 09 output. ^f Carbonyl C atom. ^g Methyl C atom. ^h The united atom topological model UAHF of the acetone molecule with hydrogen atoms summed into methyl carbon atoms.¹¹³ ⁱ Universal force field (UFF) radii.¹¹⁴ ^j Scaled UFF radii.

Table 3 shows vertical excitation energies in 23 solvents and the corresponding solvatochromic shifts relative to the gas phase calculated using the cLR/PCM/TD-M06/MG3S electrostatic model with the SMD intrinsic Coulomb radii. Table 4 presents the same quantities calculated with UFF radii scaled by 1.1. The nonelectrostatic contributions ($\Delta\omega_D$ and $\Delta\omega_H$) to the solvatochromic shifts were calculated using the model parameters D (eq 5) and H (eq 6) obtained by a least-squares fitting of $\Delta\omega$ values from eq 3, with theoretical $\Delta\omega_{EP}$ values, to the corresponding experimental values⁹ of $\Delta\omega$ over the set of 23 solvents. The optimized values of D and H are given in Table 5 for all tested models. Table 6 shows the corresponding mean signed and mean unsigned errors in $\Delta\omega$ over the set of 23 solvents.

First, we consider only electrostatic contributions $\Delta\omega_{EP}$ to the solvatochromic shifts. Tables 3 and 4 show that the electrostatic contributions ($\Delta\omega_{EP}$) depend on the values of intrinsic Coulomb radii used in electrostatic calculations, and this dependence is strongest with respect to the radius on the carbonyl oxygen. Note that the SMD radius for O is defined as a function of Abraham's hydrogen-bond acidity parameter (α) for a given solvent, whereas the SMD H and C radii and all of the UFF radii are independent of solvent. The SMD radius for O is equal to 1.52 Å for any solvent with $\alpha \geq 0.43$ (in our test set, such solvents are methanol and water), while the corresponding 1.1×UFF radius is 1.925 Å (Table 2). The smaller SMD radius on O leads to a more negative $\Delta\omega_{EP}$, as compared to $\Delta\omega_{EP}(\text{exp})$, which is more favorable for acetone in water and less favorable for acetone in methanol. There are five out of 23 solvents with $0 < \alpha < 0.43$ (more specifically, with $0 < \alpha \leq 0.11$), and there are 16 solvents with $\alpha = 0$. For these 21 solvents, the SMD oxygen radius ranges from 2.096 (for *cis*-dichloroethylene with $\alpha = 0.11$) to 2.294 Å (for *n*-pentane and others with $\alpha = 0$), and the SMD electrostatic contributions $\Delta\omega_{EP}$ to the corresponding solvatochromic shifts are similar to those obtained using the 1.1×UFF radii.

In general, the error associated with comparing the $\Delta\omega_{EP}$ values directly to experiment, averaged over 23 solvents (Table 6) decreases if a higher scaling factor for intrinsic Coulomb radii is used (e.g., compare the results for 1.3×UFF to those for UFF). However, the scaling of Coulomb radii can lead to unphysical values of the radii, and the scaling alone cannot provide a reliable method because reliable calculations must take into account not only bulk electrostatics (that is, long-range electrostatic polarization effects and the part of the short-range electrostatic effect that is adequately modeled with bulk parameters and the chosen radii) but also nonelectrostatic effects, the most significant of which are solute–solvent dispersion and hydrogen-bonding effects. On the other hand, there is no unique way to separate the bulk electrostatic contribution to the free energy of solvation from the remainder.^{88,89,115,116} Indeed, Table 5 indicates the strong dependence of optimized dispersion and hydrogen-bonding model parameters on the choice of intrinsic Coulomb radii used in bulk-electrostatic calculations, but the two-parameter model (with parameters D and H) allows one to correct the $\Delta\omega_{EP}$ contributions in the right direction relative to experiment over all sets of radii and tested solvents with only one exception, dioxane in which case the computed value of $\Delta\omega_{EP}$ (for instance, -231 with the SMD radii and -246 cm^{-1} with the 1.1×UFF radii, Tables 3 and 4) appears overly “red” with respect to the corresponding experimental value⁹ ($\Delta\omega = -305 \text{ cm}^{-1}$) because the values of the O radius in our calculations are apparently too large for dioxane, resulting in the ground electronic state of the acetone molecule in dioxane being undersolvated.

The models using the UA0 radii (in which case one assigns types to atoms and treats certain groups consisting of an atom and its covalently attached hydrogens as a united atom) are quite accurate as well. However, sometimes it is preferable to use a model that does not require the user to assign molecular-mechanics types to an atom or group, and such assignments may be ambiguous if the hybridization changes upon electronic excitation.

Table 6 indicates that the errors in VEM42/INDO/S2 calculations are smaller than those in most of the PCM/TD-DFT calculations. However, the semiempirical VEM42/INDO/S2 method is likely to underestimate the $\Delta\omega_{EP}$ contributions, especially in polar solvents, and this results in the hydrogen-bonding contribution being overestimated by eq 6. Note that INDO/S2 gives an error of $\sim 3000 \text{ cm}^{-1}$ (Table 1) for predicting the gas-phase vertical excitation energy, which is not competitive with good density functional calculations.¹¹⁷

Table 5 indicates that the sign of the H parameter is negative, and it signifies that hydrogen-bonding contributes in the direction of the shift being blue, in agreement with conclusions in earlier work.¹⁹ The sign of the D parameter is positive, meaning that solute–solvent dispersion contributes toward a red shift. Indeed, as noted in the introduction, the change in solute–solvent dispersion generally leads to red shifts. One of the tested models (Table 5) results in a negative value of D . However, this model is just a null test that does not include any electrostatics (the $\Delta\omega_{EP}$ contribu-

Table 3. Vertical Excitation Energies (ω , cm⁻¹) and Solvatochromic Shifts ($\Delta\omega$, cm⁻¹) for the $n \rightarrow \pi^*$ Transition of Acetone in 23 Solvents Calculated Using the cLR/PCM/TD-M06/MG3S Electrostatic Model with the SMD Coulomb Radii^a

name	solvent			exp		theory ^b					deviation
	ϵ	n	α	ω	$\Delta\omega$	ω_{EP}	$\Delta\omega_{EP}$	$\Delta\omega_D$	$\Delta\omega_H$	$\Delta\omega$	$\Delta\Delta\omega$
gas phase	1	1	0	35 975		36 067					
perfluoro- <i>n</i> -octane	1.7	1.3	0	36 130	-155	36 230	-164	261	0	97	252
<i>n</i> -pentane	1.8371	1.3575	0	35 950	25	36 243	-177	298	0	121	96
<i>n</i> -hexane	1.8819	1.3749	0	35 940	35	36 249	-182	308	0	126	91
<i>n</i> -heptane	1.9113	1.3878	0	35 935	40	36 253	-186	316	0	130	90
<i>n</i> -decane	1.9846	1.4102	0	35 920	55	36 263	-197	329	0	132	77
<i>n</i> -hexadecane	2.0402	1.4345	0	35 950	25	36 269	-202	342	0	140	115
carbon tetrachloride	2.228	1.4601	0	35 695	280	36 296	-229	356	0	127	-153
dioxane	2.2099	1.4224	0	36 280	-305	36 297	-231	336	0	105	410
tetraethoxysilane	2.5	1.382	0	36 104	-129	36 364	-298	312	0	15	144
diethyl ether	4.24	1.3526	0	36 155	-180	36 536	-469	295	0	-175	5
methyl acetate	6.8615	1.3614	0	36 308	-333	36 652	-586	300	0	-285	48
tetrahydrofuran	7.4257	1.405	0	36 272	-297	36 655	-589	326	0	-263	34
<i>sec</i> -butyl chloride	8.393	1.3971	0	36 132	-157	36 683	-616	321	0	-295	-138
dichloromethane	8.93	1.4242	0.1	36 300	-325	36 775	-708	337	-35	-407	-82
<i>cis</i> -dichloroethylene	9.2	1.449	0.11	36 170	-195	36 783	-716	350	-39	-405	-210
<i>tert</i> -butyl chloride	9.663	1.385	0	36 120	-145	36 713	-646	314	0	-332	-187
1,2-dichloroethane	10.125	1.4448	0.1	36 280	-305	36 792	-725	348	-35	-413	-108
acetone, neat	20.493	1.3588	0.04	36 373	-398	36 844	-778	299	-14	-493	-95
acetonitrile	35.688	1.3442	0.07	36 440	-465	36 925	-858	289	-25	-593	-128
dimethyl sulfoxide	46.826	1.417	0	36 350	-375	36 834	-767	333	0	-434	-59
propylene carbonate	62.93	1.421	0	36 393	-418	36 848	-782	335	0	-447	-29
methanol	32.613	1.3288	0.43	36 916	-941	37 535	-1,469	280	-152	-1,341	-400
water	78.3553	1.333	0.82	37 760	-1,785	37 569	-1,502	282	-289	-1,509	276

^a Solvents are sorted as in ref 9. Vertical excitation energies in the gas phase and in solution and solvatochromic shifts with respect to the gas phase are calculated over the test set of 23 solvents. Solvent descriptors are dielectric constant (ϵ), refractive index (n), and Abraham's hydrogen-bond acidity parameter (α). Theoretical solvatochromic shifts ($\Delta\omega$) are expressed in terms of the corresponding electronic-polarization (EP), dispersion (D), and hydrogen-bonding (H) components (see eqs 3–6). The $\Delta\Delta\omega$ values refer to the difference between $\Delta\omega$ (theory) and $\Delta\omega$ (experiment). ^b The PCM/TD-M06 vertical excitation energies in solution were calculated using the M06-2X/MG3S molecular geometries of acetone optimized in solution by the SMD implicit solvation model, whereas the corresponding gas-phase vertical excitation energy of acetone was calculated at the M06-2X/MG3S geometry optimized in the gas phase.

Table 4. Vertical Excitation Energies (ω , cm⁻¹) and Solvatochromic Shifts ($\Delta\omega$, cm⁻¹) for the $n \rightarrow \pi^*$ Transition of Acetone in 23 Solvents Calculated Using the cLR/PCM/TD-M06/MG3S Electrostatic Model with the UFF Coulomb Radii Scaled by the Factor of 1.1^a

name	solvent			exp		theory ^a					deviation
	ϵ	n	α	ω	$\Delta\omega$	ω_{EP}	$\Delta\omega_{EP}$	$\Delta\omega_D$	$\Delta\omega_H$	$\Delta\omega$	$\Delta\Delta\omega$
gas phase	1	1	0	35 975		36 067					
perfluoro- <i>n</i> -octane	1.7	1.3	0	36 130	-155	36 239	-173	283	0	111	266
<i>n</i> -pentane	1.8371	1.3575	0	35 950	25	36 252	-186	323	0	138	113
<i>n</i> -hexane	1.8819	1.3749	0	35 940	35	36 259	-193	335	0	142	107
<i>n</i> -heptane	1.9113	1.3878	0	35 935	40	36 263	-196	343	0	147	107
<i>n</i> -decane	1.9846	1.4102	0	35 920	55	36 275	-208	357	0	149	94
<i>n</i> -hexadecane	2.0402	1.4345	0	35 950	25	36 281	-215	372	0	157	132
carbon tetrachloride	2.228	1.4601	0	35 695	280	36 310	-244	387	0	143	-137
dioxane	2.2099	1.4224	0	36 280	-305	36 313	-246	364	0	118	423
tetraethoxysilane	2.5	1.382	0	36 104	-129	36 384	-318	339	0	22	151
diethyl ether	4.24	1.3526	0	36 155	-180	36 573	-507	320	0	-186	-6
methyl acetate	6.8615	1.3614	0	36 308	-333	36 699	-632	326	0	-306	27
tetrahydrofuran	7.4257	1.405	0	36 272	-297	36 701	-635	354	0	-281	16
<i>sec</i> -butyl chloride	8.393	1.3971	0	36 132	-157	36 731	-665	349	0	-316	-159
dichloromethane	8.93	1.4242	0.1	36 300	-325	36 697	-630	366	-169	-433	-108
<i>cis</i> -dichloroethylene	9.2	1.449	0.11	36 170	-195	36 691	-624	380	-186	-430	-235
<i>tert</i> -butyl chloride	9.663	1.385	0	36 120	-145	36 763	-696	341	0	-355	-210
1,2-dichloroethane	10.125	1.4448	0.1	36 280	-305	36 712	-645	378	-169	-436	-131
acetone, neat	20.493	1.3588	0.04	36 373	-398	36 844	-778	324	-68	-521	-123
acetonitrile	35.688	1.3442	0.07	36 440	-465	36 878	-811	314	-118	-615	-150
dimethyl sulfoxide	46.826	1.417	0	36 350	-375	36 891	-824	361	0	-463	-88
propylene carbonate	62.93	1.421	0	36 393	-418	36 905	-838	364	0	-474	-56
methanol	32.613	1.3288	0.43	36 916	-941	36 663	-596	304	-727	-1,019	-78
water	78.3553	1.333	0.82	37 760	-1,785	36 653	-586	307	-1,386	-1,665	120

^a See footnotes of Table 3.

tions are assumed to be zero), and the exception merely proves that a separate (unbalanced) treatment of bulk-electrostatic and nonelectrostatic contributions may lead to spurious results.

The SMD intrinsic Coulomb radii lead to good agreement with experiment, whereas the UFF radii need to be increased to provide good accuracy. Comparison of the $\Delta\omega_{EP}$, $\Delta\omega_D$, and $\Delta\omega_H$ components in Tables 3 and 4 indicates that the

Table 5. Model Parameters for Dispersion (D) and Hydrogen-Bonding (H) Components (cm^{-1}) Optimized over Reference Data in 23 Solvents^a

electrostatic model	radii ^b	D	H
no electrostatics ^c	n/a	-685	-2007
PCM/TD-B3LYP	Bondi	4351	-837
PCM/TD-B3LYP	SM5.42	4516	-861
PCM/TD-B3LYP	SMD	1487	-514
PCM/TD-B3LYP	UA0	1757	-1658
PCM/TD-B3LYP	UAHF	4288	-901
PCM/TD-B3LYP	UFF	2525	-1432
PCM/TD-B3LYP	1.1×UFF	1557	-1749
PCM/TD-B3LYP	1.3×UFF	362	-2111
PCM/TD-M06	Bondi	4795	-721
PCM/TD-M06	SM5.42	4909	-759
PCM/TD-M06	SMD	1655	-353
PCM/TD-M06	UA0	2035	-1588
PCM/TD-M06	UAHF	4686	-797
PCM/TD-M06	UFF	2835	-1353
PCM/TD-M06	1.1×UFF	1798	-1690
PCM/TD-M06	1.3×UFF	520	-2076
VEM42/INDO/S2	SM5.42	1589	-1646
VEM42/INDO/S2 ^d	SM5.42	3448	-1614

^aThe notation PCM refers to the cLR/PCM/TD-DFT/MG3S calculations. Unless noted otherwise, two model parameters (D and H) describing dispersion and hydrogen-bonding contributions to the solvatochromic shift for the $n \rightarrow \pi^*$ transition of acetone were optimized using eqs 3–6 over 23 experimental reference data and the corresponding electrostatic (EP) contributions calculated using the given combinations of vertical electrostatic models and Coulomb radii. Vertical excitation energies in solution were calculated using the M06-2X/MG3S molecular geometries of acetone optimized in solution by the SMD implicit solvation model, whereas the corresponding gas-phase vertical excitation energy of acetone was calculated at the M06-2X/MG3S geometry optimized in the gas phase. ^bSee footnotes of Table 2. ^cThe EP contribution is neglected. ^dTaken from ref 19, where the experimental data set used for parametrization was older and smaller than that used here.

values of $\Delta\omega_D$ calculated using the 1.1×UFF radii are close to the corresponding values calculated using the SMD radii. Indeed, the dispersion contribution in 23 solvents varies between 261 and 356 cm^{-1} when the SMD radii are used and between 283 and 387 cm^{-1} when the 1.1×UFF radii are used. However, the absolute values of $\Delta\omega_H$ in the SMD case are much smaller than those calculated using the 1.1×UFF radii. Thus, we assume that in the SMD case most of the hydrogen-bonding effect on solvatochromic shifts in solvents that exert such an effect (i.e., in solvents with nonzero Abraham's hydrogen-bond acidity parameter α) can be implicitly accounted for through the bulk-electrostatic contribution $\Delta\omega_{EP}$, in part due to the dependence of the SMD Coulomb radius of the oxygen atom on the value of α . Note that the magnitude of $\Delta\omega_{EP}$ is more sensitive to the O radius value than to the values of the Coulomb radii on C and H, which are independent of solvent in the present work. Note also that the variation in the O radius was not adopted especially for purposes of the present calculations, but has always been part of the SMD solvation model⁹⁵ where it is required to obtain accurate equilibrium free energies of solvation for ground-state ions in nonaqueous solutions. The absolute values of the H parameter obtained using the SMD radii for bulk electrostatics are relatively small, and they are 3–5 times smaller than those optimized within the 1.1×UFF model (Table 5). Upon neglect of the $\Delta\omega_H$ contribution (H

= 0), the RMSE over 23 solvents increases only slightly if the SMD radii are used (for instance, in the case of TD-M06 calculations from 175 to 189 cm^{-1}), but the error increase is more significant (from 158 to 369 cm^{-1}) if the 1.1×UFF radii are used. We conclude that the use of UFF radii scaled by 1.1 (Gaussian 09 default radii) for calculation of the vertical $n \rightarrow \pi^*$ excitation energy of acetone in water (and, perhaps, in other cases) requires the $\Delta\omega_H$ correction for the bulk-electrostatic (electron-polarization) component because the bulk-electrostatic contribution is underestimated when the 1.1×UFF radii are used. We note, however, that our assessment at this point is specific to the case we are studying here, and by extension perhaps carbonyls in general, but will need to be further investigated for other chromophores.

Table 6 also shows the errors in $\Delta\omega$ using the gas-phase molecular geometry of acetone in vertical excitation energy calculations in solution instead of geometries optimized for the ground state in the particular solvent using the SMD⁹⁵ solvation model. Note that the use of solvent-specific geometries of acetone is the default in the present study, while in the previous study¹⁹ only a gas-phase geometry was used. Optimizing the molecular geometry of acetone in solution leads to a substantial increase of the C–O distance in more polar media, which, in turn, leads to a substantial decrease of the $\Delta\omega_{EP}$ contribution to the corresponding solvatochromic shift. For example, the M06-2X/MG3S optimization yields $R_c(\text{CO}) = 1.204 \text{ \AA}$ in the gas phase and $R_c(\text{CO}) = 1.218 \text{ \AA}$ in water. The corresponding TD-M06/MG3S gas-phase excitation energy at the gas-phase geometry is 36 067 cm^{-1} , the cLR/PCM/TD-M06/MG3S (with the SMD Coulomb radii) excitation energy of acetone in water at the water-specific optimized geometry is 37 569 cm^{-1} , and the corresponding EP contribution to the shift is $\Delta\omega_{EP} = -1502 \text{ cm}^{-1}$. Using the TD-M06/MG3S gas-phase geometry in the cLR/PCM/TD-M06/MG3S calculation leads to an excitation energy of acetone in water equal to -38 026 cm^{-1} , with the corresponding EP contribution equal to $\Delta\omega_{EP} = -1959 \text{ cm}^{-1}$. This trend agrees with the results of previous calculations.^{22,28} Table 6 shows that using the gas-phase geometry decreases the accuracy of the theoretical values of $\Delta\omega$ across all of the electrostatic models tested in the present work.

Table 7 presents partial atomic charges and dipole moments for the ground and for the first excited electronic state of the acetone molecule in the gas phase and in *n*-hexane and water. The partial atomic charges were calculated using the Merz–Singh–Kollman electrostatic potential fitting method^{118,119} implemented in Gaussian 09.⁹³ The dipole moment of the acetone molecule in solution increases in comparison with the gas-phase dipole moment in both the ground and the first excited electronic states, and the polarity of the acetone molecule increases with the polarity of the solvent (as seen by comparing the results for *n*-hexane and water). The ground-state dipole moment of the acetone molecule is higher than the excited-state dipole moment, and the difference is larger for more polar solvents because the acetone molecule is more favorably solvated in its ground state than in the nascent excited state. The trends in the partial atomic charges on the carbon and oxygen atoms of the

Table 6. Errors in Vertical Excitation Energy Solvatochromic Shifts (cm^{-1}) for the $n \rightarrow \pi^*$ Transition of Acetone in 23 Solvents^{a,b}

electrostatic model	radii ^e	gas-phase geometry ^c				liquid-phase geometry ^d			
		EP		EPDH		EP		EPDH	
		MSE	MUE	MSE	MUE	MSE	MUE	MSE	MUE
no electrostatics ^f	n/a	280 ^g	320 ^g	4	137	280 ^g	320 ^g	4	137
PCM/TD-B3LYP	Bondi	-963	963	2	397	-767	786	2	352
PCM/TD-B3LYP	SM5.42	-982	982	2	381	-797	815	2	341
PCM/TD-B3LYP	SMD	-461	461	2	172	-244	289	2	130
PCM/TD-B3LYP	UA0	-435	491	3	185	-212	350	3	138
PCM/TD-B3LYP	UAHF	-944	944	2	376	-751	774	2	333
PCM/TD-B3LYP	UFF	-590	624	2	240	-375	460	3	190
PCM/TD-B3LYP	1.1×UFF	-393	456	3	165	-167	322	3	119
PCM/TD-B3LYP	1.3×UFF	-152	280	3	97	88	203	4	93
PCM/TD-M06	Bondi	-1053	1053	2	430	-861	867	2	387
PCM/TD-M06	SM5.42	-1061	1061	2	409	-880	886	2	371
PCM/TD-M06	SMD	-504	504	2	184	-288	319	2	140
PCM/TD-M06	UA0	-492	540	3	203	-270	390	3	154
PCM/TD-M06	UAHF	-1025	1025	2	406	-834	846	2	364
PCM/TD-M06	UFF	-653	678	2	263	-440	510	3	211
PCM/TD-M06	1.1×UFF	-443	499	3	181	-217	357	3	132
PCM/TD-M06	1.3×UFF	-185	303	3	103	55	197	4	88
VEM42/INDO/S2	SM5.42	-264	358	3	112	-181	318	3	100

^a Mean signed (MSE) and mean unsigned (MUE) errors refer to the difference between theoretical and experimental solvatochromic shifts. EPDH denotes electronic (E), polarization (P), dispersion (D), and hydrogen-bonding (H) contributions to the solvatochromic shift. The EP contributions were calculated using the given combinations of vertical electrostatic models and Coulomb radii, whereas the corresponding DH contributions were either neglected (the EP column) or determined within the two-parameter model (eqs 3–6). ^b The corresponding experimental energies include 23 reference data from ref 9. ^c Vertical excitation energies in solution and in the gas phase were calculated using the ground-state gas-phase geometry optimized at the M06-2X/MG3S level of theory. ^d Vertical excitation energies in solution were calculated using the M06-2X/MG3S molecular geometries of acetone optimized in solution by the SMD implicit solvation model, whereas the corresponding gas-phase vertical excitation energy of acetone was calculated at the M06-2X/MG3S geometry optimized in the gas phase. The corresponding model parameters D and H used to calculate nonelectrostatic (DH) contributions are given in Table 5. ^e See footnotes of Table 2. ^f The EP contribution is neglected. ^g The solvatochromic shift (EPDH) is assumed to be zero.

carbonyl group and in the magnitudes of the dipole moment of the acetone molecule in *n*-hexane and water obtained using the SMD Coulomb radii are qualitatively similar to the trends in the corresponding values obtained using the UFF radii scaled by 1.1 (Table 7). However, the use of the SMD Coulomb radii results in a higher dipole moment in the more polar solvent. It is difficult to further quantify the accuracy of either the SMD or the 1.1×UFF radii for predicting the dipole moment in solution because of the absence of experimental liquid-phase dipole moments. Note that experimental dipole moments are readily available for gas-phase molecules¹²⁰ but not for molecules in solution, where the dipole moment is not even uniquely defined, irrespective of the electronic state.

Table 8 contains predicted solvatochromic shifts on the vertical $n \rightarrow \pi^*$ excitation transition of acetone for 18 common solvents that were not included in the training set of 23 solvents.⁹ These solvents are a subset of the Minnesota Solvent Descriptor Database.¹⁰³ The Supporting Information gives predictions of solvatochromic shifts for all 178 solvents of the database.

3.2. Mixed Discrete-Continuum Models. We have studied the effect of adding up to 12 water molecules explicitly to the acetone molecule with the solute and explicit water molecules treated as a supermolecule immersed in an aqueous continuum. The clusters were constructed using 20 random acetone–water configurations obtained from a molecular dynamics (MD) simulation. The MD simulation was performed with the MacroModel utility¹²¹ in the Maestro Version 8.5 computational package¹²² using a cubic box of

$37.3 \times 37.3 \times 37.3 \text{ \AA}^3$ containing 1733 water molecules and one rigid acetone molecule (fixed at the aqueous-phase geometry optimized at the SMD/M06-2X/MG3S level of theory), the OPLS_2005 force field¹²³ with the force-field-defined electrostatic treatment and charges, a temperature of 300 K, a time step of 1.5 fs, and an equilibration time of 200 ps. The simulation was carried out for 2 ns, and configurations were written every 100 ps. The OPLS_2005 force field¹²³ is an enhanced version of the OPLS all-atom force field of Jorgensen et al.¹²⁴ For each of the 20 configurations, we defined a supermolecular cluster by retaining the 12 water molecules closest to the oxygen atom in the acetone molecule (as determined by the distance from the carbonyl oxygen to the nearest hydrogen of the water molecule). We also defined smaller clusters as subsets of these in which we retained only the 1, 2, 5, 8, or 10 closest water molecules.

Table 9 shows values of the vertical excitation energy (ω) and solvatochromic shift ($\Delta\omega$) for the $n \rightarrow \pi^*$ transition of acetone in water estimated by averaging over the corresponding energies calculated for the 20 acetone–water clusters including n water molecules explicitly (where n varies between 1 and 12) and treating the rest within the continuum approximation. All of the energies in Table 9 nominally contain only bulk-electrostatic (electron-polarization) effects with $H = D = 0$, but it should be noticed that hydrogen bonding and dispersion-like attractive noncovalent interactions between acetone and the explicit water molecules are included by treating acetone and the explicit waters with the M06 density functional.

Table 7. Partial Atomic Charges (au) and Dipole Moments (debye) Calculated for the Acetone Molecule in the Ground and First Excited Electronic States^a

method	SMD radii ^b		1.1×UFF radii ^b		gas ^c	
	ground	excited	ground	excited	ground	excited
Solvent: <i>n</i> -Hexane						
<i>q</i> (C)	0.80	0.43	0.80	0.43	0.79	0.43
<i>q</i> (O)	-0.59	-0.34	-0.59	-0.34	-0.56	-0.32
μ_q	3.45	2.00	3.45	1.99	3.14	1.78
μ_p	3.47	2.02	3.46	2.01	3.15	1.80
Solvent: Water						
<i>q</i> (C)	0.86	0.46	0.81	0.44	0.78	0.43
<i>q</i> (O)	-0.71	-0.40	-0.65	-0.37	-0.57	-0.32
μ_q	4.67	2.72	4.16	2.37	3.23	1.78
μ_p	4.70	2.76	4.19	2.40	3.24	1.78

^a The notations *q*(C) and *q*(O) refer to the partial atomic charges on the carbon and the oxygen atom in the carbonyl group, respectively. The notations μ_q and μ_p refer to the dipole moment calculated based on the partial atomic charges and the quantum-mechanical electronic density, respectively. The partial atomic charges were obtained with the Merz–Singh–Kollman electrostatic potential fitting method.^{118,119} The ground- and excited-state electronic densities were calculated at the TD-M06/MG3S level of theory. The excited-state electronic density of the acetone molecule in solution was calculated using the ground-state reaction field. ^b Calculated using the polarizable continuum solvent approximation and the corresponding Coulomb radii at the SMD/M06-2X/MG3S ground-state geometry of the acetone molecule optimized in *n*-hexane and water. ^c Calculated in the gas phase at the SMD/M06-2X/MG3S ground-state geometry of the acetone molecule optimized in *n*-hexane and water; the calculation at the corresponding gas-phase optimized geometry was also performed: *q*(C) = 0.79, *q*(O) = -0.56, μ_q = 3.12, μ_p = 3.13 (ground state), and *q*(C) = 0.43, *q*(O) = -0.32, μ_q = 1.79, μ_p = 1.80 (excited state). For comparison, the experimental gas-phase value of μ is 2.88 ± 0.03 debye (ground state).¹²⁰

Table 8. Predicted Solvatochromic Shifts ($\Delta\omega$, cm⁻¹) for the $n \rightarrow \pi^*$ Transition of Acetone in 18 Common Organic Solvents^a

name	ϵ	<i>n</i>	α	$\Delta\omega$
acetic acid	6.2528	1.3720	0.61	-934
benzene	2.2706	1.5011	0	147
carbon disulfide	2.6105	1.6319	0	177
chloroform	4.7113	1.4459	0.15	-292
cyclohexane	2.0165	1.4266	0	136
diethylamine	3.5766	1.3864	0.08	-171
diiodomethane	5.3200	1.7425	0.05	-26
<i>N,N</i> -dimethylformamide	37.2190	1.4305	0	-417
ethanol	24.852	1.3611	0.37	-1135
ethyl acetate	5.9867	1.3723	0	-250
methylcyclohexane	2.0240	1.4231	0	132
4-methyl-2-pentanone	12.8870	1.3962	0	-360
2-methyl-1-propanol	16.7770	1.3955	0.37	-1034
1-octanol	9.8629	1.4295	0.37	-876
pyridine	12.9780	1.5095	0	-271
tetrachloroethene	2.2680	1.5053	0	148
toluene	2.3741	1.4961	0	126
2,2,2-trifluoroethanol	26.7260	1.2907	0.57	-1414

^a Solvatochromic shifts are calculated using the cLR/PCM/TD-M06/MG3S method and SMD Coulomb radii for vertical bulk-electrostatic energies augmented with the corresponding solute–solvent dispersion and hydrogen-bonding contributions. Solvent descriptors are dielectric constant (ϵ), refractive index (*n*), and Abraham's hydrogen-bond acidity parameter (α).¹⁰³

Figure 1 shows an example of the cluster containing five explicit water molecules. The value of $\Delta\omega$ averaged over 20 clusters (with *n* water molecules in each) depends on *n*,

Table 9. Vertical Excitation Energies (ω , cm⁻¹) and Solvatochromic Shifts ($\Delta\omega$, cm⁻¹) for the $n \rightarrow \pi^*$ Transition of Acetone in Water Calculated Using Acetone–Water Clusters and Implicit Solvent Models^a

method	radii	<i>n</i>	ω	$\Delta\omega$	SD
gas					
TD-B3LYP ^b			36 147		
TD-M06 ^b			36 067		
INDO/S2 ^b			33 055		
experiment (ref 9)			35 975		
continuum ^c					
PCM/TD-B3LYP	SMD	0	37 510	-1363	
PCM/TD-B3LYP	SMD	12	37 631	-1484	479
PCM/TD-M06	SMD	0	37 569	-1502	
PCM/TD-M06	SMD	1	37 796	-1730	313
PCM/TD-M06	SMD	2	37 863	-1797	452
PCM/TD-M06	SMD	5	37 788	-1721	466
PCM/TD-M06	SMD	8	37 769	-1703	463
PCM/TD-M06	SMD	10	37 763	-1697	492
PCM/TD-M06	SMD	12	37 742	-1675	474
PCM/TD-M06 ^d	SMD	1	38 373	-2307	
PCM/TD-M06 ^d	SMD	2	39 003	-2936	
PCM/TD-M06	1.1×UFF	0	36 653	-586	
PCM/TD-M06	1.1×UFF	1	37 089	-1023	406
PCM/TD-M06	1.1×UFF	12	37 212	-1146	675
VEM42/INDO/S2	SM5.42	0	33 661	-606	
VEM42/INDO/S2	SM5.42	12	33 222	-167	256
VEM42/INDO/S2 ^e	SM5.42	12	33 262	-207	237
experiment (ref 9)			37 760	-1785	

^a The MG3S basis set was used in all TD-DFT calculations.

^b The vertical excitation energy of bare acetone in the gas phase was calculated with a given method at the M06-2X/MG3S optimized gas-phase geometry. ^c Vertical excitation energies of acetone–water clusters (ω) containing *n* water molecules and that of bare acetone (*n* = 0) in water were calculated using given implicit solvation models and Coulomb radii. The values of ω were averaged over 20 random acetone–water configurations obtained from an MD simulation unless noted otherwise. They contain no dispersion and hydrogen-bonding corrections, and they nominally correspond to ω_{EP} . The $\Delta\omega$ values were calculated relative to the gas-phase vertical excitation energy of bare acetone. SD stands for standard deviation. ^d Not an average, but a single-point energy calculation with the SMD/M06-2X/MG3S optimized liquid-phase geometry of an acetone–water cluster corresponding to the global minimum (Figure 2). ^e Calculated over 100 random acetone–water configurations obtained from an MD simulation.

but this dependence is small for *n* ≥ 5. The averaged theoretical values of $\Delta\omega$ are in good agreement with the experimental value⁹ (-1785 cm⁻¹) when we use the SMD intrinsic Coulomb radii: -1675 cm⁻¹ (M06, *n* = 12). The use of the 1.1×UFF radii instead leads to an underestimated blue shift (-1146 cm⁻¹) in these calculations (M06, *n* = 12). The least accurate model in these calculations is VEM42/INDO/S2, which severely underestimates the blue shift (-167 cm⁻¹) with respect to experiment, apparently due to an underestimate of the hydrogen bonding in the acetone–water clusters at the INDO/S2 level.

Table 9 also shows a result for the VEM42/INDO/S2 model when we average over all 100 acetone–water clusters. The absolute value of ω changes by 40 cm⁻¹ in comparison with the value of ω averaged over only 20 clusters. Although the averaging over larger numbers of MD snapshots can lead to a more precise prediction of the solvatochromic shift for a given model chemistry, we have chosen to use only 20 acetone–water clusters in most of the calculations because

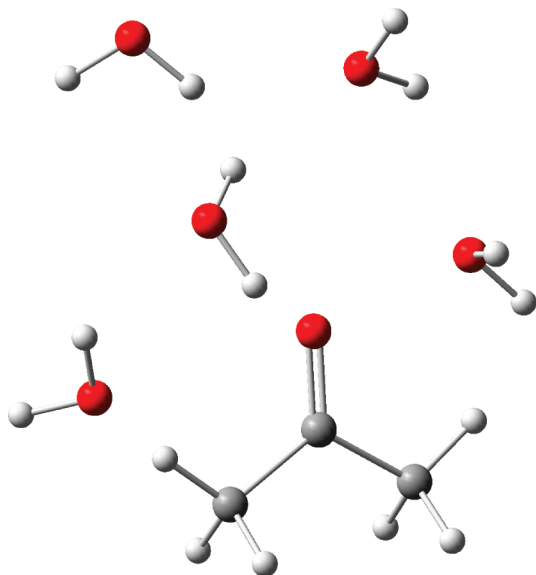


Figure 1. An example of the solute–solvent cluster containing five explicit water molecules. Hydrogen atoms are white, carbon is gray, and oxygen is red.

this choice provides for a reasonable balance of computational time-saving and small sampling error.

The values of ω and $\Delta\omega$ averaged over the 20 MD clusters in aqueous solution are also compared in Table 9 to those calculated for a single solute–solvent cluster with one or two water molecules at the ground-state molecular geometry optimized in solution by the SMD/M06-2X/MG3S method (Figure 2). This approach with a single solute–solvent cluster, which is widely used in the literature, overestimates the blue shift by 522–1151 cm^{-1} with respect to the experimental value,⁹ and in the present study it is substantially less accurate than the method that considers a dynamically generated distribution of solvent orientations.

3.3. Comparison with Other Models. Solvatochromic shifts on the vertical $n \rightarrow \pi^*$ excitation of acetone have been the subject of numerous theoretical studies.^{11–33,75} Here, we discuss results of several recent calculations of the vertical $n \rightarrow \pi^*$ excitation energy of acetone in water.^{21,22,28,75} The results of these previous studies^{21,22,28,75} are presented in Table 10, where they are compared to selected results obtained in the present work.

Cossi and Barone²¹ employed the conductor-like version of PCM (C-PCM) that incorporated a computation of the final (excited) state at the CASSCF/6-31G(d) level. Quantum-mechanical terms for solute–solvent dispersion and repulsion interactions were included in the solute Hamiltonian to take into account the nonelectrostatic contributions to solvatochromic shifts, assuming no contribution from cavitation effects.²¹ They obtained the following dispersion–repulsion contributions to solvatochromic shifts:²¹ 606, 409, 397, and 378 cm^{-1} in cyclohexane, dichloromethane, ethanol, and water, respectively (within the sign convention adopted in the present work). Except for cyclohexane, these numbers are in qualitative agreement with the values of $\Delta\omega_D$ predicted in the present work using the cLR/PCM/TD-M06/MG3S method with the SMD Coulomb radii, 338, 337, 300, and 282 cm^{-1} in the same four solvents, respectively. Aquilante

et al.²² carried out a TD-DFT computation of the acetone ultraviolet spectrum in aqueous solution using the PBE0 density functional and mixed discrete/continuum models. The authors noticed that the attachment of two explicit water molecules to the carbonyl oxygen of the acetone molecule (with the cluster being optimized) led to an overly blue solvatochromic shift predicted by PCM, and the computed shifts were sensitive to the orientation of the solvent molecules around the carbonyl group.²² These observations by Aquilante et al.²² are confirmed in the present study with M06.

Aidas et al.²⁸ applied the combined linear response coupled cluster/molecular mechanics (CC/MM) scheme including mutual polarization effects in the coupling Hamiltonian to the study of the vertical $n \rightarrow \pi^*$ excitation energy of acetone in water, by averaging over 800 solute–solvent configurations obtained from a molecular dynamics simulation. A spherical cutoff radius of 10 Å was applied to retain 125–148 nearby water molecules, which were treated at the MM level using SPC and SPCpol potentials (the latter accounts for explicit solvent polarization effects), while the acetone molecule was treated quantum mechanically by the CCSD/aug-cc-pVDZ method.²⁸ In our computation of the vertical $n \rightarrow \pi^*$ excitation energy of acetone in water using solute–solvent clusters, we include a much smaller number of water molecules explicitly (up to 12 water molecules) because we treat them quantum mechanically (rather than by molecular mechanics), and we treat the rest within the continuum approximation (cLR/PCM), which Aidas et al.²⁸ did not use. In additional calculations, Aidas et al. treated two explicit water molecules quantum mechanically by the CCSD/aug-cc-pVDZ method with no implicit solvent.²⁸

Chipman⁷⁵ calculated the vertical $n \rightarrow \pi^*$ excitation energy of acetone in water using an extension of the SVPE and SS(V)PE solvation models to treat the nonequilibrium solvation effects. The SVPE and SS(V)PE results were compared to those obtained using the conductor-like and dielectric versions of PCM, respectively, C-PCM and D-PCM. The author discussed the quantum mechanical computation of vertical transitions in solution in regard to volume polarization effects arising from penetration of the solute charge density outside the cavity.⁷⁵ The sensitivity of the computed vertical energies to the cavity size was noted.⁷⁵

4. Summary

The inclusion of solvent effects beyond bulk electrostatics, for example, cavitation and dispersion, has been shown to be a key ingredient in calculating accurate free energies of solvation for equilibrated solutes in their ground electronic states.^{88,89,115,116} A related challenge is the inclusion of these same effects in the treatment of condensed-phase electronic excitation. In this work, we computed the vertical $n \rightarrow \pi^*$ electronic excitation energy of acetone in 22 nonaqueous solvents and in water including the contributions of (i) nonequilibrium electrostatic polarization, (ii) changes in solvent–solute dispersion, (iii) changes in solvent–solute hydrogen-bonding, and (iv) ground-state geometry relaxation. The solvatochromic shifts ($\Delta\omega$) on the vertical electronic excitation energies relative to the gas phase were computed

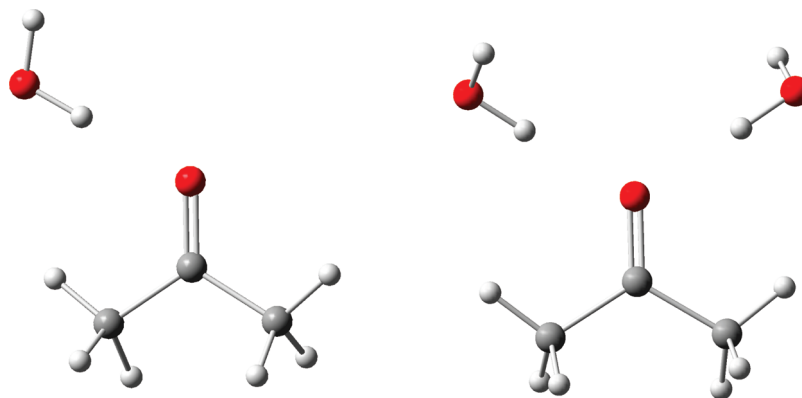


Figure 2. Molecular structures of the solute–solvent clusters containing one and two explicit water molecules optimized in solution at the SMD/M06-2X/MG3S level of theory. See caption to Figure 1 for atom colors.

Table 10. Recent Computations on the Vertical $n \rightarrow \pi^*$ Excitation Energies (ω , cm^{-1}) of Acetone in Water

method	reference	$\omega(\text{gas})$	$\omega(\text{water})$	$\Delta\omega$
PCM/CASSCF/6-31G(d) ^{a,b}	Cossi and Barone ²¹	37 195	37 964	−769
PCM/CASSCF/6-31G(d)/cluster ^{a,c,d}	Cossi and Barone ²¹	37 195	39 592	−2397
PCM/TD-PBE0/6-311++G(2d,2p) ^b	Aquilante et al. ²²	35 730	36 779	−1049
PCM/TD-PBE0/6-311++G(2d,2p)/cluster ^c	Aquilante et al. ²²	35 730	38 795	−3065
PCM/TD-PBE0/6-311++G(2d,2p)/cluster ^{c,e}	Aquilante et al. ²²	35 730	37 263	−1533
CCSD/aug-cc-pVDZ/cluster ^f	Aidas et al. ²⁸	36 698	38 279	−1581
QM/MM(SPCpol) CCSD/aug-cc-pVDZ ^g	Aidas et al. ²⁸	36 698	37 795	−1097
SS(V)PE/CASSCF/6-311(2+)G(d) ^h	Chipman ⁷⁵	36 618	37 910	−1292
SS(V)PE/CASSCF/6-311(2+)G(d) ⁱ	Chipman ⁷⁵	36 618	38 876	−2258
cLR/PCM/TD-M06 ^{b,j}	present	36 067	37 576	−1509
cLR/PCM/TD-M06/cluster ^k	present	36 067	37 742	−1675
experiment ^l	Bayliss et al. ⁶	36 100	37 800	−1700
experiment ^l	Renge ⁹	35 975	37 760	−1785 ± 7

^a With nonequilibrium electrostatics and quantum-mechanical dispersion–repulsion contributions. ^b No explicit water molecules. ^c With two explicit water molecules. ^d Using a cluster called “Conformation 1” in ref 21. ^e Averaged over 40 acetone–water clusters (see ref 22 for more detail). ^f Linear response coupled cluster theory with two explicit water molecules added and with no implicit solvent. ^g The combined linear response coupled cluster/molecular mechanics (CC/MM) scheme including mutual polarization effects; the acetone molecule was treated at the CCSD/aug-cc-pVDZ level, whereas at most 148 explicit water molecules were treated at the MM level. ^h With the cavity size defined by solute electronic isodensity contours of $\rho_0 = 0.0005$ au. ⁱ With the cavity size defined by solute electronic isodensity contours of $\rho_0 = 0.002$ au. ^j See Table 3. ^k Averaged over 20 acetone–water clusters containing 12 explicit water molecules treated as part of a quantum mechanical supermolecule (Table 9). ^l UV absorption.

using two different continuum solvation models for calculating the electrostatic component of the shifts. One of the electrostatic models is based on the generalized Born approximation, and another solves the nonhomogeneous Poisson equation for electrostatics in terms of the integral-equation-formalism polarizable continuum model. Both models use a two-response-time electrostatic reaction field corresponding to equilibrated fast (electronic) solvent response and nonequilibrated Franck–Condon-defined slow (nuclear) solvent response. The dependence of calculated solvatochromic shifts on the values of the intrinsic atomic Coulomb radii used for construction of the boundary between the solute cavity and the solvent continuum in the bulk electrostatic calculations was investigated.

In addition to studies using a fully continuum representation of water, we also examined the effect of adding up to 12 explicit water molecules for the calculation of $\Delta\omega$ for acetone in aqueous solution. Sets of 20 random acetone–water configurations were taken from molecular dynamics trajectories, and in each case solvatochromic shifts were computed for the cluster of solute and explicit solvent molecules surrounded by a continuum representing the rest of the aqueous solvent.

On the basis of the results of these calculations, we draw five key conclusions:

(1) Although nonbulk-electrostatic contributions to solvatochromic shifts are usually neglected in modern calculations, they are not necessarily small in magnitude. In particular, we find that reasonable choices of cavity radii lead to dispersion contributions to the solvatochromism of acetone on the order of 300–400 cm^{-1} in all solvents.

(2) Hydrogen-bonding contributions were even larger in magnitude than dispersion effects in solvents having strong hydrogen-bond-donating ability. When the electrostatic cavity is constructed from oxygen radii that are large in water and independent of solvent (e.g., scaled UFF radii), explicit corrections for hydrogen bonding must be made to improve predictive accuracy. Some SMD radii, including oxygen radii, by contrast depend on the solvent hydrogen-bond donating character, and as a result additional corrections for the effect of changes in hydrogen bonding on acetone solvatochromism were up to an order of magnitude smaller.

(3) When nonbulk-electrostatic contributions are explicitly included in the solvation model, they can ameliorate unphysical choices for the atomic radii used in the construction of the solute cavity in the purely continuum electrostatics

calculation. Nevertheless, analysis of trends in the magnitudes (and signs) of the correction terms renders clear which sets of radii should be regarded as most physically appropriate. From such an analysis, we conclude that SMD radii, which were optimized for the computation of ground-state free energies of solvation and not previously examined for the prediction of solvatochromism, represent the optimal available combination of accuracy and physicality. The use of UFF radii scaled by a factor of 1.1 also gave good results for the quantitative prediction of solvatochromic shifts, but only after the addition of a substantially larger correction for changes in hydrogen bonding than was needed for SMD. As noted above, this reflects the ability of the SMD model to capture such effects through the use of solvent-dependent atomic cavity radii. Insofar as post hoc corrections for changes in hydrogen bonding are not easily calculated (in the absence of a large set of solvatochromic training data), and moreover such corrections are not included in the relaxation of the excited-state wave function (which may lead to other computed properties being inaccurate), we recommend the use of SMD radii in place of scaled UFF radii. We note that this conclusion is based on the limited data in the present Article, and its generality needs to be investigated beyond carbonyl compounds, but the calculations presented here indicate that significant quantitative errors can be obtained if solvatochromic shifts are calculated with scaled or unscaled molecular mechanics radii *and only the influence of bulk electrostatics is considered*.

(4) The optimization of acetone–water clusters including a small number of water molecules does not lead to substantially improved predictions for the aqueous solvatochromic shift when these clusters are embedded in the continuum. It appears that the water–acetone hydrogen-bonding interactions in the optimized cluster are too strong as compared to those that would be found in bulk solution. Indeed, when clusters are not optimized but instead removed from molecular dynamics trajectories in explicit water, the influence of bulk water molecules on the first solvation shell around acetone manifests itself in smaller predicted blue shifts that are in much better agreement with experiment. Predicted shifts from the extracted-cluster protocol converge very quickly with the number of water molecules n chosen to be explicitly retained. With SMD radii, results for $n = 2, 5, 8, 10,$ and 12 are within one standard deviation of one another when averaged over 20 snapshots. For $n = 12$, $\Delta\omega = -1675 \text{ cm}^{-1}$ calculated with the M06 density functional approximation agrees well with the experimental⁹ $\Delta\omega = -1785 \text{ cm}^{-1}$. The use of UFF radii scaled by a factor of 1.1 for the same clusters leads to an underestimated blue shift, $\Delta\omega = -1146 \text{ cm}^{-1}$. As the acetone carbonyl oxygen is effectively buried by the explicit solvent shell, this suggests that the improved performance of the SMD radii is associated not only with the acetone molecule, but also the surrounding water molecules. We conclude that treating only a few water molecules explicitly with structures optimized for the cluster (supersolute) (rather than averaging over a canonical ensemble of structures selected from a simulation of the bulk) leads to significant systematic errors.

(5) Because the standard deviation of the solvatochromic shift predicted from the molecular dynamics trajectory snapshots measures the broadening of the excited-state absorption due to thermal fluctuations in the surrounding medium, we were able to see that the width of the distribution of solvatochromic shifts in water is large. In particular, we calculate a standard deviation of 474 cm^{-1} of the single-molecule shifts from their mean. If one compares this to the predicted solvatochromic shift of -1675 cm^{-1} , one sees that the width of the distribution is not insignificant, and one should keep this heterogeneity in mind in interpreting solvatochromic shifts in general.

Finally, in addition to considering experimental results previously reported in the literature, we have also predicted the vertical $n \rightarrow \pi^*$ electronic excitation energies of acetone in 160 nonaqueous solvents that were not included in the training set of 23 solvents⁹ (see the Supporting Information). These predictions should prove interesting for comparison to future measurements.

Acknowledgment. We are grateful to Dr. Carlos P. Sosa (IBM and Minnesota Supercomputing Institute) for invaluable assistance. This work was supported by the Army Research Office under grant US ARMY RES LAB/W911NF09-1-0377 and by the National Science Foundation under grants CHE06-10183, CHE07-04974, and CHE09-56776.

Supporting Information Available: Vertical excitation energies and solvatochromic shifts for the $n \rightarrow \pi^*$ transition of acetone in water calculated using acetone–water clusters and implicit solvent models; vertical excitation energies and solvatochromic shifts for the $n \rightarrow \pi^*$ transition of acetone in water calculated using acetone–water clusters with one and two explicit water molecules; vertical excitation energies and solvatochromic shifts for the $n \rightarrow \pi^*$ transition of acetone in the 178 nonaqueous solvents in the Minnesota Solvent Descriptor Database (the set of 23 solvents studied in the paper includes, in addition to water, 18 nonaqueous solvents that are among 178 solvents in that database and 4 nonaqueous solvents (perfluoro-*n*-octane, tetraethoxysilane, *tert*-butyl chloride, and propylene carbonate) that are not in the database); and Cartesian coordinates of the acetone molecule in the ground electronic state optimized in the gas phase and in solution. This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- (1) Bayliss, N. S.; McRae, E. G. *J. Phys. Chem.* **1954**, *58*, 1006.
- (2) Pimentel, G. C. *J. Am. Chem. Soc.* **1957**, *79*, 3323.
- (3) Ito, M.; Inuzuka, K.; Imanishi, S. *J. Am. Chem. Soc.* **1960**, *82*, 1317.
- (4) Balasubramanian, A.; Rao, C. N. R. *Spectrochim. Acta* **1962**, *18*, 1337.
- (5) Hayes, W. P.; Timmons, C. J. *Spectrochim. Acta* **1965**, *21*, 529.
- (6) Bayliss, N. S.; Wills-Johnson, G. *Spectrochim. Acta* **1968**, *24A*, 551.

- (7) Yudasaka, M.; Hosoya, H. *Bull. Chem. Soc. Jpn.* **1978**, *51*, 1708.
- (8) Xu, H.; Wentworth, P. J.; Howell, N. W.; Joens, J. A. *Spectrochim. Acta* **1993**, *49A*, 1171.
- (9) Renge, I. *J. Phys. Chem. A* **2009**, *113*, 10678.
- (10) Fox, T.; Rösch, N. *Chem. Phys. Lett.* **1992**, *191*, 33.
- (11) Pappalardo, R. R.; Reguero, M.; Robb, M. A.; Frisch, M. *Chem. Phys. Lett.* **1993**, *212*, 12.
- (12) Rösch, N.; Zerner, M. C. *J. Phys. Chem.* **1994**, *98*, 5817.
- (13) Ten-no, S.; Hirata, F.; Kato, S. *J. Chem. Phys.* **1994**, *100*, 7443.
- (14) Gao, J. *J. Am. Chem. Soc.* **1994**, *116*, 9324.
- (15) Liao, D. W.; Mebel, A. M.; Chen, Y.-T.; Lin, S.-H. *J. Phys. Chem. A* **1997**, *101*, 9925.
- (16) Serrano-Andrés, L.; Fülischer, M. P.; Karlström, G. *Int. J. Quantum Chem.* **1997**, *65*, 167.
- (17) Mennucci, B.; Cammi, R.; Tomasi, J. *J. Chem. Phys.* **1998**, *109*, 2798.
- (18) Coutinho, K.; Saavedra, N.; Canuto, S. *THEOCHEM* **1999**, *466*, 69.
- (19) Li, J.; Cramer, C. J.; Truhlar, D. G. *Int. J. Quantum Chem.* **2000**, *77*, 264.
- (20) Martin, M. E.; Sanchez, M. L.; Olivares del Valle, F. J.; Aguilar, M. A. *J. Chem. Phys.* **2000**, *113*, 6308.
- (21) Cossi, M.; Barone, V. *J. Chem. Phys.* **2000**, *112*, 2427.
- (22) Aquilante, F.; Cossi, M.; Crescenzi, O.; Scalmani, G.; Barone, V. *Mol. Phys.* **2003**, *101*, 1945.
- (23) Röhrig, U. F.; Frank, I.; Hutter, J.; Laio, A.; VandeVondele, J.; Rothlisberger, U. *ChemPhysChem* **2003**, *4*, 1177.
- (24) Bernasconi, L.; Sprik, M.; Hutter, J. *J. Chem. Phys.* **2003**, *119*, 12417.
- (25) Coutinho, K.; Canuto, S. *THEOCHEM* **2003**, *632*, 235.
- (26) Crescenzi, O.; Pavone, M.; De Angelis, F.; Barone, V. *J. Phys. Chem. B* **2005**, *109*, 445.
- (27) Sulpizi, M.; Röhrig, U. F.; Hutter, J.; Rothlisberger, U. *Int. J. Quantum Chem.* **2005**, *101*, 671.
- (28) Aidas, K.; Kongsted, J.; Osted, A.; Mikkelsen, K. V.; Christiansen, O. *J. Phys. Chem. A* **2005**, *109*, 8001.
- (29) Öhrn, A.; Karlström, G. *Theor. Chem. Acc.* **2007**, *117*, 441.
- (30) Fonseca, T. L.; Coutinho, K.; Canuto, S. *J. Chem. Phys.* **2007**, *126*, 034508.
- (31) Minezawa, N.; Kato, S. *J. Chem. Phys.* **2007**, *126*, 054511.
- (32) Lin, Y.-l.; Gao, J. *J. Chem. Theory Comput.* **2007**, *3*, 1484.
- (33) Gomes, A. S. P.; Jacob, C. R.; Visscher, L. *Phys. Chem. Chem. Phys.* **2008**, *10*, 5353.
- (34) Liptay, W. *Z. Naturforsch.* **1965**, *20A*, 1441.
- (35) Aguilar, M. A.; Olivares del Valle, F. J.; Tomasi, J. *J. Chem. Phys.* **1993**, *98*, 7375.
- (36) Ridley, J. E.; Zerner, M. C. *Theor. Chim. Acta* **1973**, *32*, 111.
- (37) Zerner, M. C. *Rev. Comput. Chem.* **1991**, *2*, 313.
- (38) Li, J.; Williams, B.; Cramer, C. J.; Truhlar, D. G. *J. Chem. Phys.* **1999**, *110*, 724.
- (39) Zerner, M. C.; Ridley, J. E.; Bacon, A. D.; Edwards, W. D.; Head, J. D.; McKelvey, J.; Culberson, J. C.; Knappe, P.; Cory, M. G.; Weiner, B.; Baker, J. D.; Parkinson, W. A.; Kannis, D.; Yu, J.; Roesch, N.; Kotzian, M.; Tamm, T.; Karelson, M. M.; Zheng, X.; Pearl, G.; Broo, A.; Albert, K.; Cullen, J. M.; Cramer, C. J.; Truhlar, D. G.; Li, J.; Hawkins, G. D.; Liotard, D. A. *ZINDO computer program - version 99.1*, 1999.
- (40) Cossi, M.; Barone, V. *J. Chem. Phys.* **2001**, *115*, 4708.
- (41) Runge, E.; Gross, E. K. U. *Phys. Rev. Lett.* **1984**, *52*, 997.
- (42) Casida, M. E. In *Time-Dependent Density-Functional Response Theory for Molecules*; Chong, D. P., Ed.; World Scientific: Singapore, 1995; Vol. 1, p 155.
- (43) Bauernschmitt, R.; Ahlrichs, R. *Chem. Phys. Lett.* **1996**, *256*, 454.
- (44) Stratmann, R. E.; Scuseria, G. E.; Frisch, M. J. *J. Chem. Phys.* **1998**, *109*, 8218.
- (45) Tomasi, J.; Persico, M. *Chem. Rev.* **1994**, *94*, 2027.
- (46) Cramer, C. J.; Truhlar, D. G. *Chem. Rev.* **1999**, *99*, 2161.
- (47) Cramer, C. J.; Truhlar, D. G. In *Free Energy Calculations in Rational Drug Design*; Reddy, M. R., Erion, M. D., Eds.; Kluwer Academic/Plenum: New York, 2001; p 63.
- (48) Tomasi, J.; Mennucci, B.; Cammi, R. *Chem. Rev.* **2005**, *105*, 2999.
- (49) Mennucci, B. In *Continuum Solvation Models in Chemical Physics*; Mennucci, B., Cammi, R., Eds.; Wiley: Chichester, U.K., 2007; p 110.
- (50) Marcus, R. A. *J. Chem. Phys.* **1956**, *24*, 966.
- (51) Aguilar, M. A. *J. Phys. Chem. A* **2001**, *105*, 10393.
- (52) Marcus, R. A. *J. Chem. Phys.* **1956**, *24*, 979.
- (53) Wolynes, P. G. *J. Chem. Phys.* **1987**, *86*, 5133.
- (54) Hsu, C.-P.; Song, X.; Marcus, R. A. *J. Phys. Chem. B* **1997**, *101*, 2546.
- (55) Basilevsky, M. V.; Parsons, D. F.; Vener, M. V. *J. Chem. Phys.* **1998**, *108*, 1103.
- (56) Caricato, M.; Ingrosso, F.; Mennucci, B.; Tomasi, J. *J. Chem. Phys.* **2005**, *122*, 154501.
- (57) Hoijtink, G. J.; de Boer, E.; van der Meij, P. H.; Weijland, W. P. *Recl. Trav. Chim. Pays-Bas Belg.* **1956**, *75*, 487.
- (58) Peradejordi, F. *Cahiers Phys.* **1963**, *17*, 393.
- (59) Klopman, G. *Chem. Phys. Lett.* **1967**, *1*, 200.
- (60) Tapia, O. In *Quantum Theory of Chemical Reactions*; Daudel, R., Pullman, A., Salem, L., Viellard, A., Eds.; Wiley: London, 1981; Vol. 2, p 25.
- (61) Tucker, S. C.; Truhlar, D. G. *Chem. Phys. Lett.* **1989**, *157*, 164.
- (62) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. *J. Am. Chem. Soc.* **1990**, *112*, 6127.
- (63) Mikkelsen, K. V.; Cesar, A.; Ågren, H.; Jensen, H. J. Aa. *J. Chem. Phys.* **1995**, *103*, 9010.
- (64) Miertuš, S.; Scrocco, E.; Tomasi, J. *Chem. Phys.* **1981**, *55*, 117.
- (65) Miertuš, S.; Tomasi, J. *Chem. Phys.* **1982**, *65*, 239.
- (66) Improta, R.; Barone, V.; Scalmani, G.; Frisch, M. J. *J. Chem. Phys.* **2006**, *125*, 054103.

- (67) Improta, R.; Scalmani, G.; Frisch, M. J.; Barone, V. *J. Chem. Phys.* **2007**, *127*, 074504.
- (68) Scalmani, G.; Frisch, M. J.; Mennucci, B.; Tomasi, J.; Cammi, R.; Barone, V. *J. Chem. Phys.* **2006**, *124*, 094107.
- (69) Caricato, M.; Mennucci, B.; Tomasi, J.; Ingrosso, F.; Cammi, R.; Corni, S.; Scalmani, G. *J. Chem. Phys.* **2006**, *124*, 124520.
- (70) Cancès, E.; Mennucci, B.; Tomasi, J. *J. Chem. Phys.* **1997**, *107*, 3032.
- (71) Mennucci, B.; Tomasi, J. *J. Chem. Phys.* **1997**, *106*, 5151.
- (72) Mennucci, B.; Cancès, E.; Tomasi, J. *J. Phys. Chem. B* **1997**, *101*, 10506.
- (73) Tomasi, J.; Mennucci, B.; Cancès, E. *J. Mol. Struct. (THEOCHEM)* **1999**, *464*, 211.
- (74) Chipman, D. M. *J. Chem. Phys.* **2009**, *131*, 014103.
- (75) Chipman, D. M. *J. Chem. Phys.* **2009**, *131*, 014104.
- (76) Zhan, C.-G.; Bentley, J.; Chipman, D. M. *J. Chem. Phys.* **1998**, *108*, 177.
- (77) Chipman, D. M. *Theor. Chem. Acc.* **2002**, *107*, 80.
- (78) Cancès, E.; Mennucci, B. *J. Chem. Phys.* **2001**, *114*, 4744.
- (79) Cammi, R.; Corni, S.; Mennucci, B.; Tomasi, J. *J. Chem. Phys.* **2005**, *122*, 104513.
- (80) Corni, S.; Cammi, R.; Mennucci, B.; Tomasi, J. *J. Chem. Phys.* **2005**, *123*, 134512.
- (81) Li, J.; Zhu, T.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. A* **2000**, *104*, 2178.
- (82) Pekar, S. I. *Untersuchungen über die Elektronentheorie der Kristalle*; Akademie-Verlag: Berlin, Germany, 1954.
- (83) Cramer, C. J.; Truhlar, D. G. In *Solvent Effects and Chemical Reactivity*; Tapia, O., Bertran, J., Eds.; Understanding Chemical Reactivity Series; Kluwer: Dordrecht, The Netherlands, 1996; Vol. 17, p 1.
- (84) Zhu, T.; Li, J.; Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. *J. Chem. Phys.* **1998**, *109*, 9117. errata: **1999**, *111*, 5624 and **2000**, *113*, 3930.
- (85) Li, J.; Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. *Chem. Phys. Lett.* **1998**, *288*, 293.
- (86) Li, J.; Zhu, T.; Hawkins, G. D.; Winget, P.; Liotard, D. A.; Cramer, C. J.; Truhlar, D. G. *Theor. Chem. Acc.* **1999**, *103*, 9.
- (87) Cramer, C. J.; Truhlar, D. G. In *Trends and Perspectives in Modern Computational Science*; Maroulis, G., Simos, T. E., Eds.; Lecture Series on Computer and Computational Sciences 6; Brill/VSP: Leiden, The Netherlands, 2006; p 112.
- (88) Cramer, C. J.; Truhlar, D. G. *Acc. Chem. Res.* **2008**, *41*, 760.
- (89) Cramer, C. J.; Truhlar, D. G. *Acc. Chem. Res.* **2009**, *42*, 493.
- (90) Li, J.; Zhu, T.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. A* **1998**, *102*, 1820.
- (91) Zerner, M. C.; Ridley, J. E.; Bacon, A. D.; Edwards, W. D.; Head, J. D.; McKelvey, J.; Culberson, J. C.; Knappe, P.; Cory, M. G.; Weiner, B.; Baker, J. D.; Parkinson, W. A.; Kannis, D.; Yu, J.; Roesch, N.; Kotzian, M.; Tamm, T.; Karelson, M. M.; Zheng, X.; Pearl, G.; Broo, A.; Albert, K.; Cullen, J. M.; Li, J.; Hawkins, G. D.; Thompson, J. D.; Kelly, C. P.; Liotard, D. A.; Cramer, C. J.; Truhlar, D. G. *ZINDO-MN1.2, Quantum Theory Project*; University of Florida, Gainesville, and Department of Chemistry, University of Minnesota, Minneapolis, 2005.
- (92) Lynch, B. J.; Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2003**, *107*, 1384.
- (93) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, N. J.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, Ö.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. *Gaussian 09*, revision A.02; Gaussian, Inc.: Wallingford, CT, 2009.
- (94) *Gaussian09 User's Reference: SCRF*; http://www.gaussian.com/g-tech/g_ur/k_scrf.htm (accessed May 19, 2010).
- (95) Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. B* **2009**, *113*, 6378.
- (96) Zhao, Y.; Truhlar, D. G. *Theor. Chem. Acc.* **2008**, *120*, 215.
- (97) Zhao, Y.; Truhlar, D. G. *Acc. Chem. Res.* **2008**, *41*, 157.
- (98) Reichardt, C. *Solvents and Solvent Effects in Organic Chemistry*, 2nd ed.; VCH: Weinheim, 1990; Chapter 6.
- (99) Abraham, M. H.; Grellier, P. L.; Prior, D. V.; Duce, P. P.; Morris, J. J.; Taylor, P. J. *J. Chem. Soc., Perkin Trans. 2* **1989**, 699.
- (100) Abraham, M. H. *Chem. Soc. Rev.* **1993**, *22*, 73.
- (101) Abraham, M. H. *J. Phys. Org. Chem.* **1993**, *6*, 660.
- (102) Abraham, M. H. In *Quantitative Treatment of Solute/Solvent Interactions; Theoretical and Computational Chemistry Series Vol. 1*; Politzer, P., Murray, J. S., Eds.; Elsevier: Amsterdam, 1994; p 83.
- (103) Winget, P.; Dolney, D. M.; Giesen, D. J.; Cramer, C. J.; Truhlar, D. G. *Minnesota Solvent Descriptor Database version 1999*; University of Minnesota: Minneapolis, MN, 1999; <http://comp.chem.umn.edu/solvation/mnsddb.pdf> (accessed March 18, 2010).
- (104) Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098.
- (105) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785.
- (106) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648.
- (107) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. *J. Phys. Chem.* **1994**, *98*, 11623.
- (108) Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2006**, *110*, 13126.
- (109) Zhao, Y.; Truhlar, D. G. *J. Chem. Phys.* **2006**, *125*, 194101.
- (110) Foresman, J. B.; Head-Gordon, M.; Pople, J. A.; Frisch, M. J. *J. Phys. Chem.* **1992**, *96*, 135.
- (111) Bondi, A. *J. Phys. Chem.* **1964**, *68*, 441.
- (112) Barone, V.; Improta, R.; Rega, N. *Theor. Chem. Acc.* **2004**, *111*, 237.

- (113) Barone, V.; Cossi, M.; Tomasi, J. *J. Chem. Phys.* **1997**, *107*, 3210.
- (114) Rappé, A. K.; Casewit, C. J.; Colwell, K. S.; Goddard, W. A., III; Skiff, W. M. *J. Am. Chem. Soc.* **1992**, *114*, 10024.
- (115) Curutchet, C.; Cramer, C. J.; Truhlar, D. G.; Ruiz-López, M. F.; Rinaldi, D.; Orozco, M.; Luque, F. J. *J. Comput. Chem.* **2003**, *24*, 284.
- (116) Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. *J. Chem. Theory Comput.* **2008**, *4*, 877.
- (117) Jacquemin, D.; Perpète, E. A.; Ciofini, I.; Adamo, C.; Valero, R.; Zhao, Y.; Truhlar, D. G. *J. Chem. Theory Comput.* **2010**, *6*, 2071.
- (118) Singh, U. C.; Kollman, P. A. *J. Comput. Chem.* **1984**, *5*, 129.
- (119) Besler, B. H.; Merz, K. M., Jr.; Kollman, P. A. *J. Comput. Chem.* **1990**, *11*, 431.
- (120) *CRC Handbook of Chemistry and Physics*; Lide, D. R., Ed.; Taylor and Francis: Boca Raton, FL, 2010; Vol. 90 (Internet Version 2010, <http://www.hbcnetbase.com>).
- (121) *MacroModel, version 9.6*; Schrödinger, LLC: New York, 2008.
- (122) *Maestro Version 8.5.111, MMshare Version 1.7.110*; Schrödinger, LLC: New York, 2008.
- (123) 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.
- (124) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1996**, *118*, 11225.

CT100267S

Understanding the Resonance Raman Scattering of Donor–Acceptor Complexes using Long-Range Corrected DFT

Daniel W. Silverstein and Lasse Jensen*

*Department of Chemistry, The Pennsylvania State University, 104 Chemistry Building,
University Park, Pennsylvania 16802*

Received May 26, 2010

Abstract: The optical properties involving charge-transfer states of the donor–acceptor electron-transfer complexes carbazole/tetracyanoethylene (carbazole/TCNE) and hexamethylbenzene/tetracyanoethylene (HMB/TCNE) were investigated by utilizing the time-dependent theory of Heller to simulate absorbance and resonance Raman spectra. Excited-state properties were obtained using time-dependent density functional theory (TDDFT) using the global hybrid B3LYP and the long-range corrected LC- ω PBE functionals and compared with experimental results. It is shown that, while reasonable simulations of the absorbance spectra can be made using B3LYP, the resonance Raman spectra for both complexes are poorly described. The LC- ω PBE functional gives a more accurate representation of the excited-state potential energy surfaces in the Franck–Condon region for charge-transfer states, as indicated by the good agreement with the experimental resonance Raman spectrum. For the carbazole/TCNE complex, which includes contributions from two overlapping excited states on its absorbance spectrum, interference effects are discussed, and it is found that detuning from resonance with an excited state results in interference along with other factors. Total vibrational reorganization energy for both complexes is discussed, and it is found that both B3LYP and LC- ω PBE yield reasonable estimates of this quantity compared with experiment.

Introduction

Electron-transfer reactions are fundamental processes involved in chemistry and biology. Some important applications of these processes include studies of photocatalysis,^{1,2} charge transfer (CT) in polymers,³ and numerous other biological and chemical processes.^{4–6} Resulting from those applications, investigations involving the rate of electron transfer and optimization of that process are a major research focus.

For molecules that undergo a change in oxidation state due to electron transfer, the molecular geometry changes as well. The rate of electron transfer between molecules strongly depends on molecular geometries of the donor and acceptor molecules before and after the electron-transfer event.^{7–9} A more quantitative analysis of structural changes due to

electron-transfer requires knowledge of which normal modes are Franck–Condon (FC) active, because it is these modes that are strongly influenced by electron transfer. Using resonance Raman spectroscopy allows the CT state being investigated to be characterized and the information about the rate of electron transfer to be quantified.^{7,8,10–12}

Several methods for modeling resonance Raman scattering have been developed. In the vibronic theory of Albrecht et al.,^{13–15} the Born–Oppenheimer (BO) approximation is used to separate each vibronic state into a product of the electronic and vibrational wave functions. Then, the transition dipole moment is expanded as a Taylor series in the nuclear coordinates. This allows the Raman transition polarizability to be represented as a sum of terms, where the first is the A term (FC) and the second is the B term (Herzberg–Teller). Another formulation is the time-dependent formalism developed by Heller et al.^{16–21} This method uses wave packet dynamics to describe the time-dependent overlap of the final

* Corresponding author. E-mail: jensen@chem.psu.edu.

state and the initial vibrational wave function that propagates along the excited-state potential energy surface. Unlike the vibronic theory, the time-dependent method avoids the computationally demanding summation over intermediate vibrational states.^{16,19,22} Both methods can be used with harmonic approximations of the ground- and excited-state potential energy surfaces that are displaced along the normal coordinate, in the independent mode displaced harmonic oscillator (IMDHO) method.²¹ Excited-state displacements calculated using the IMDHO method are proportional to the excited-state gradient at the ground-state equilibrium geometry, which can be used to model resonance Raman spectra and also be applied to studying electron-transfer rates using Marcus theory.^{7,8,10}

Density functional theory (DFT) has been applied in several cases for studying vibronic structure of molecules in resonance Raman scattering.^{23–25} Traditional DFT exchange–correlation (XC) functionals have been shown to largely underestimate the CT excitations of weakly interacting systems,^{26–29} but this effect has been demonstrated to be partially corrected using the long-range corrected (LC) DFT.^{30–33} Although many studies have shown the usefulness of LC-DFT for correctly describing CT excitation energies,^{30–33} there has not been a study of resonance Raman scattering of CT excited states using the LC functionals. This will provide an additional validation of the LC functionals since it is crucial to accurately describe the curvature of the excited-state surface for correctly modeling resonance Raman scattering.

Here we will present a detailed study of the resonance Raman scattering of two donor–acceptor complexes using LC-DFT combined with Heller’s time-dependent theory of Raman scattering. Time-dependent density functional theory (TDDFT) is used to evaluate excited-state displacements within the IMDHO method for the hexamethylbenzene/tetracyanoethylene (HMB/TCNE) and carbazole/tetracyanoethylene (carbazole/TCNE) complexes. The absorption and resonance Raman scattering spectra are then simulated using the time-dependent theory. Complexes similar to those studied in this work have been investigated using the LC-functional BNL (Baer–Neuhauser–Livshits)^{33,34} to investigate how the attenuation parameter in that functional can be tuned to give improved descriptions of excitation energies for CT states in comparison to experiment.³⁵ For both of these systems, their resonance Raman scattering have been measured experimentally,^{36–38} thus enabling a comprehensive comparison between theory and experiment. The low-energy portion of the optical absorbance spectra for these complexes include excitations where an electron is transferred from the donor (carbazole or HMB) to the acceptor (TCNE) molecule.^{36–38} The complex HMB/TCNE has a single CT state on its absorbance spectrum, while carbazole/TCNE has two energetically close CT states that overlap in one band on the absorbance spectrum. For the carbazole/TCNE complex, contributions from two CT states result in interference effects, which have been observed in previous studies that focused on experimentally derived fits to resonance Raman spectra.^{37,39,40} Total vibrational reorganization ener-

gies are also compared between the XC functionals and experiment for both complexes.

Theory

Expressions for the absorbance cross-section (σ_a) and Raman polarizability ($\alpha_{\rho\lambda}^n$) can be obtained by applying both the FC and BO approximations. In the time-dependent formalism, expressions for σ_a and $\alpha_{\rho\lambda}^n$ can be rewritten as respective full- and half-Fourier transforms, resulting in^{16–21}

$$\sigma_a = \frac{4\pi}{3\hbar c} E_L \sum_n (\mu^{0n})^2 \times \text{Re} \int_0^\infty \langle i | i_n(t) \rangle e^{i(E_L + \nu_{i0})t - \Gamma_n t - (1/2)\Theta^2 t^2} dt \quad (1)$$

and

$$\alpha_{\rho\lambda}^n = \sum_n \mu_\rho^{0n} \mu_\lambda^{n0} \times i \int_0^\infty \langle f | i_n(t) \rangle e^{i(E_L + \nu_{i0})t - \Gamma_n t - (1/2)\Theta^2 t^2} dt \quad (2)$$

For both expressions, E_L defines the energy of the incident radiation, n defines the electronic state (where 0 is the electronic ground state), μ^{0n} defines the electronic transition dipole moment for an excitation between electronic states 0 and n , ν_{i0} is the energy of vibrational state $|i\rangle$, and $|i_n(t)\rangle = e^{-i\hat{H}_n t} |i\rangle$ is the wavepacket corresponding to the time-dependent nuclear wave function of electronic state n . The homogeneous broadening for this system is treated phenomenologically with the addition of Γ_n , which allows for each excited state to have a different lifetime. It is often difficult to fit absorbance spectra with only the homogeneous broadening parameter Γ_n because the Lorentzian line shape resulting from $\exp(-\Gamma t)$ decays too slowly, which is found to worsen the fit on the red edge of the absorption spectrum.²⁰ This can be compensated by including an inhomogeneous broadening parameter, Θ , where both eqs 1 and 2 represent convolutions of Gaussian and Lorentzian line shapes (Voigt line shape). In particular for the Raman polarizability^{41,42} given in eq 2, the final vibrational state $|f\rangle$ is involved in the dynamics, and the subscripts ρ and λ refer to x , y , and z directions of the transition dipole moment vectors and polarizability tensor elements.

In order to calculate the overlaps between different vibrational states, the IMDHO model^{16,20,21} was used. This method relies on assumptions that the ground- and excited-state potential energy surfaces are harmonic and have the same normal-mode composition and frequencies. The excited-state displacement relative to the ground-state equilibrium position of the potential is given by the shift Δ_k^n in dimensionless normal coordinates. Use of the IMDHO method allows for the overlap integrals to be written

$$\langle i | i_n(t) \rangle = e^{-\sum_j s_j^n (1 - e^{-i\omega_j t}) - i(\nu_{i0} - E_{0n})t} \quad (3)$$

for determining the absorption cross-section and

$$\langle f | i_n(t) \rangle = \prod_k \left\{ \frac{(-1)^{m_k} (\Delta_k^n)^{m_k}}{(2^{m_k} m_k!)^{1/2}} (1 - e^{-i\omega_k t})^{m_k} \right\} \times e^{-\sum_j s_j^n (1 - e^{-i\omega_j t}) - i(\nu_{i0} - E_{0n})t} \quad (4)$$

for the Raman polarizability. For these expressions, $s_k^n = (\Delta_k^n)^2/2$ are the Huang–Rhys factors, and m_k is the excitation number for the k th normal mode of vibrational state $|f\rangle$.

After determining the Raman polarizabilities, the differential Raman scattering cross-section can be calculated to compare with experimental measurements.^{43–45} For measurements of scattered radiation 90° from the direction of propagation of the incident radiation, the differential Raman scattering cross-section is given by

$$\frac{d\sigma}{d\Omega} = \frac{\pi^2}{\varepsilon_0^2} (\nu_{in} - \nu_{k0})^4 \left(\frac{45a_k^2 + 7\gamma_k^2}{45} \right) \times \frac{1}{1 - \exp[-hc\nu_{k0}/k_B T]} \quad (5)$$

where ν_{in} is the energy of the incident radiation and T is the temperature (assumed to be 300 K in the present work). The tensor invariants a_k and γ_k are the isotropic and anisotropic polarizability averages, given by

$$a_k = \frac{1}{3} \{ (\alpha_{xx})_k + (\alpha_{yy})_k + (\alpha_{zz})_k \} \quad (6)$$

and

$$\gamma_k^2 = \frac{1}{2} \{ [(\alpha_{xx})_k - (\alpha_{yy})_k]^2 + [(\alpha_{yy})_k - (\alpha_{zz})_k]^2 + [(\alpha_{zz})_k - (\alpha_{xx})_k]^2 + 6[(\alpha_{xy})_k - (\alpha_{yz})_k + (\alpha_{zx})_k]^2 \} \quad (7)$$

In order to perform the integrals for obtaining the absorbance cross-section and Raman polarizability, the dimensionless excited-state displacements (Δ_k^n) must be calculated. When the potential energy surface is assumed to be harmonic,²¹ Δ_k^n relates to the partial derivative of the excited-state electronic energy with respect to a ground-state normal mode at the ground-state equilibrium position:

$$\left(\frac{\partial E^n}{\partial q_k} \right)_{q_k=0} = -\nu_{k0} \Delta_k^n \quad (8)$$

The excited-state electronic energy gradients in eq 8 are calculated using a three-point central differences formula around the ground-state equilibrium geometry. For convenience, the derivatives are initially determined in mass-weighted normal coordinates, Q_k , not the dimensionless normal coordinates q_k . However, it is easy to convert between the two using the relationship:⁴⁶

$$\frac{\partial E^n}{\partial q_k} = \left(\frac{\partial Q_k}{\partial q_k} \right) \frac{\partial E^n}{\partial Q_k} = \sqrt{\frac{\hbar}{2\pi c \nu_{k0}}} \frac{\partial E^n}{\partial Q_k} \quad (9)$$

Excited-state electronic energy gradients in terms of mass-weighted normal coordinates were evaluated using formulas similar to those presented by Reiher et al. for numerical derivatives of the polarizability tensor elements,⁴⁷ replacing the polarizability with electronic energy at positions displaced along each normal mode.

Electron-transfer rates for harmonic potential energy surfaces can be quantified if the energy penalty for transferring an electron, i.e. the energy difference between the charge-separated state at the equilibrium geometry of both

the neutral system and the charge-separated state, is known. This energy penalty is called the reorganization energy, λ_{tot} .⁹ Generally, λ_{tot} is partitioned into two components, the solvent reorganization energy (λ_s) and vibrational reorganization energy (λ_v), as

$$\lambda_{\text{tot}} = \lambda_s + \lambda_v \quad (10)$$

The solvent reorganization energy is the energy cost resulting from solvent molecules reorienting themselves after the electron transfer takes place, in order to optimize the solvent–complex interactions. Vibrational reorganization energy results from changes in the molecular geometry due to electron transfer. For a harmonic free energy surface, the total vibrational reorganization energy can be written as a sum of single-mode contributions

$$\lambda_v = \frac{1}{2} \sum_k (\Delta_k^n)^2 \nu_{k0} = \sum_k \lambda_{v,k} \quad (11)$$

Analysis of the single-mode contributions to the vibrational reorganization energy compared with what is found experimentally yields complementary information to examining the features of resonance Raman spectra.

Computational Details

In the long-range corrected (LC) approach, the interelectronic repulsion is partitioned into separate short- and long-range terms, which is given for electronic separation r_{12} as

$$\frac{1}{r_{12}} = \frac{1 - \text{erf}(\omega r_{12})}{r_{12}} + \frac{\text{erf}(\omega r_{12})}{r_{12}} \quad (12)$$

where ω is the attenuation parameter, the first term on the right side of eq 12 is for the short-range part of the exchange, and the second term of the right side of that equation is used for the long-range part of the exchange. Recently, we implemented several LC functionals into NWChem⁴⁸ based on the general approach produced by Hirao and co-workers³⁰ for constructing the short-range generalized gradient approximation (GGA) XC functional.^{49,50} An alternative procedure based on a model for the Perdew–Burke–Ernzerhof (PBE) exchange hole has been presented by Scuseria and co-workers.^{51,52} LC functionals based on this procedure, LC- ω PBE and LC- ω PBEh, have been shown to lead to a good description of both ground- and excited-state properties.⁵³ Here we have implemented these functionals into NWChem.⁴⁸

The ground-state equilibrium geometry and normal modes for the HMB/TCNE and carbazole/TCNE complexes were determined using the B3LYP functional⁵⁴ and 6-31G* basis set. Normal-mode frequencies were scaled by 0.98 from the B3LYP values to obtain better agreement with experimental frequencies. Optical properties, including excited-state energies used for determining the dimensionless displacements, were calculated using the TDDFT linear response method^{55,56} in NWChem. The dimensionless displacements were calculated based on the B3LYP structure, and the normal modes using either B3LYP or LC- ω PBE were used for calculating the excitation energies. For the LC- ω PBE functional, ω was

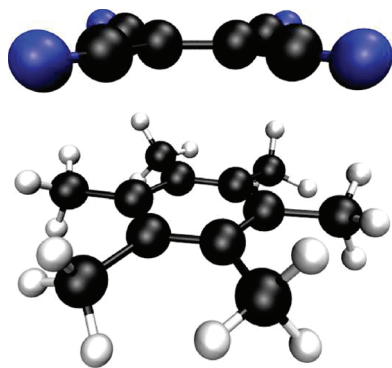


Figure 1. Optimized structure for the HMB/TCNE complex using B3LYP/6-31G*.

chosen to be $0.30a_0^{-1}$. This value is based on previous applications where it was shown that this value leads to the best performance for excited-state properties, such as excitation energies.^{53,57} Excitation energies using the LC- ω PBEh functional yielded very similar results to the LC- ω PBE functional and are thus not reported.

Simulation parameters for the absorbance spectrum were determined by shifting the peak position and changing the peak width to match the experimental absorbance spectrum using a mixture of homogeneous and inhomogeneous broadening. Peak height was normalized to match the experimental absorbance spectrum by applying a scale factor to the DFT results. Repositioned excitation energies and homogeneous and inhomogeneous broadening parameters for each functional were then used to simulate the resonance Raman spectra. The scale factors used to normalize the absorbance maxima were not applied to the resonance Raman spectra. Additional details of the fitting procedure are described in the Supporting Information.

Results and Discussion

HMB/TCNE. The structure for the HMB/TCNE complex is shown in Figure 1. In this system, the electron-donor molecule is HMB, and the electron acceptor is TCNE. Experimentally, the absorbance maximum is found to be located at 532 nm (2.33 eV) with a transition dipole moment estimated at 1.644 au.³⁷ The LC- ω PBE functional places the excitation at 499 nm (2.48 eV) with a transition dipole moment of 1.218 au, while B3LYP finds the excitation to

be at 668 nm (1.86 eV) with a transition dipole moment of 1.597 au. LC- ω PBE locates the excited state closer to where it is found experimentally than B3LYP. Discrepancies between the experimental and the theoretical excitation energies may be attributed to solvent effects that are not included in the TDDFT calculations and the basis set dependence of the excitation energies.

Simulated absorbance spectra for the HMB/TCNE complex obtained using B3LYP and LC- ω PBE are shown in Figure 2. Also, shown in Figure 2 is the simulated absorbance spectrum using data fitted to the experimental spectrum taken from ref 37. For this complex, one symmetric peak is observed with an absorbance maximum positioned at 532 nm that corresponds to an intermolecular charge-transfer state between the HMB donor and TCNE acceptor. Calculations using LC- ω PBE find that the first excitation is the highest occupied molecular orbital (HOMO)-1 to lowest unoccupied molecular orbital (LUMO) transition, which is a dark state, while the second excitation is the HOMO to LUMO transition that is a bright state. In this case, both the HOMO and HOMO-1 come from HMB, and the LUMO is from TCNE. B3LYP yields a similar character for the frontier molecular orbitals, however, it reorders the energy and the intensity of the two excitations.

Examination of Figure 2 shows that each method gives a reasonably accurate model of the experimental data. It is coincidental that each method uses the same homogeneous broadening parameter Γ , but the inhomogeneous broadening is dominant due to its much larger magnitude. The variation in the Δ_k^n values is reflected in part by the inhomogeneous broadening parameter Θ which varies from 1000 cm^{-1} for the experimental fitted data to 1800 cm^{-1} for the B3LYP functional. Also, it is apparent that each description differs because each fit is scaled by a different factor. While the Δ_k^n values from B3LYP and experiment cause the absorbance cross-section to be overestimated by about 25%, using Δ_k^n values from LC- ω PBE results in the underestimation of the absorbance cross-section by about the same factor. The scale factor that is applied to the experimental fit, in this case, is likely due to the different model for the inhomogeneous broadening used in this work compared to the original work of Myers et al.³⁷

Resonance Raman spectra simulated at 530 nm with the optimum modeling parameters for the absorbance spectra are

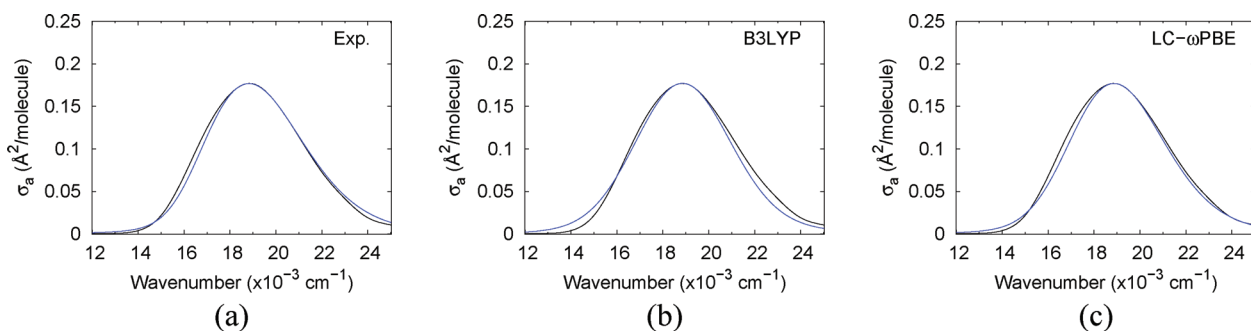


Figure 2. Absorbance spectra for HMB/TCNE with fits (blue line) compared to the experimental absorbance spectrum (black line, ref 37). The fit to experimental data (labeled "Exp.") used $\Gamma = 300\text{ cm}^{-1}$, $\Theta = 1000\text{ cm}^{-1}$, and scale factor = 0.725. Using the B3LYP functional: $\Gamma = 300\text{ cm}^{-1}$, $\Theta = 1800\text{ cm}^{-1}$, and scale factor = 0.761. For LC- ω PBE: $\Gamma = 300\text{ cm}^{-1}$, $\Theta = 1400\text{ cm}^{-1}$, and scale factor = 1.29.

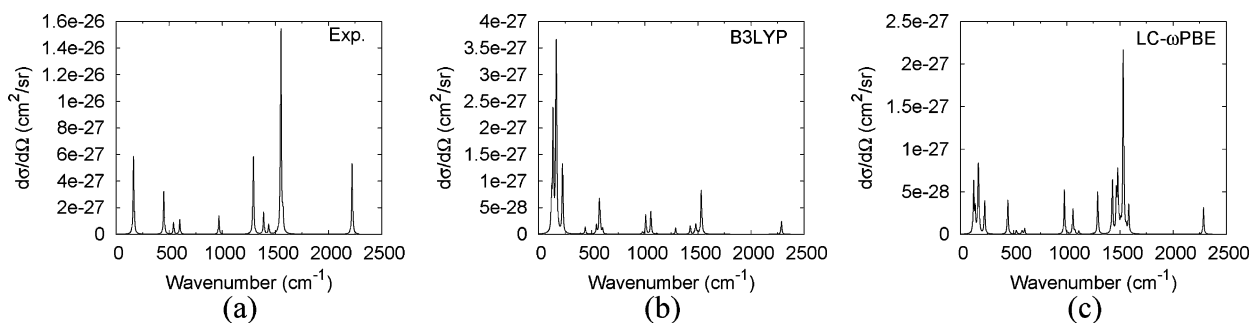


Figure 3. Resonance Raman spectra at 530 nm for HMB/TCNE using the optimum modeling parameters for the absorbance spectra plotted in Figure 2. Data for the spectrum labeled “Exp.” are from ref 37. Spectra are broadened with a Lorentzian function with 10 cm^{-1} width.

presented in Figure 3 using B3LYP and LC- ω PBE. For comparison we also plot the simulated Resonance Raman spectrum using the data obtained by fitting to the experimental spectrum, which we will refer to as the experimental spectrum since it is essentially identical to what was found by Myers et al.³⁷ A qualitative comparison of the B3LYP spectrum with the experimental resonance Raman spectrum indicates that B3LYP significantly underestimates the relative peak intensities of the high-energy normal modes. Comparing the data from the LC- ω PBE functional with the experimental resonance Raman spectrum shows that LC- ω PBE yields much better agreement with experiment. For both experiment and LC- ω PBE, the most intense peak results from the C=C stretch at 1570 cm^{-1} on the TCNE structure, whereas for B3LYP it involves the intermolecular donor–acceptor (D–A) stretching motion at 165 cm^{-1} (see Supporting Information for complete normal mode assignments).

There are some noticeable differences between the spectrum from LC- ω PBE and the experimental spectrum. The region around 1000 cm^{-1} has only one important mode (due primarily to a C–CH₃ stretch and CH₃ deformation on HMB) in the experimental spectrum, but LC- ω PBE indicates there are two modes with similar mode structure and intensity at 956.85 and 1056.58 cm^{-1} . Also, LC- ω PBE overestimates the intensities of several modes between 1460 and 1500 cm^{-1} relative to the large peak at 1531.11 cm^{-1} . It is also found that the lower frequency region (below 250 cm^{-1}) has some additional features, but these low-frequency modes are not very accurately described in the harmonic approximation. Inclusion of anharmonic effects in the model would be a method to test if anharmonicity is important, but this is not feasible for large molecular systems like the HMB/TCNE complex. There may also be an improvement in the description of these modes if solvent effects are included in the TDDFT calculations, which have been shown to be important for describing intramolecular CT states of rhodamine 6G.²⁵ It may also be important to include dispersion corrections to the XC functional in complexes, such as HMB/TCNE, like those used in the DFT-D method.^{58,59}

Quantitative inspection of the peak intensities of the different resonance Raman spectra show that the agreement between theory and experiment is not perfect, but it is interesting that changing between the two XC functionals yields such a dramatic change in relative peak intensities. The resonance Raman spectrum simulated with LC- ω PBE

yields differential Raman scattering cross-sections that differ by a range of factors between 4 and 10 when compared with those derived experimentally. A similar comparison between the spectrum from B3LYP and that from experiment shows that, resulting from the poor description of the high-energy normal modes, the differential Raman scattering cross-sections vary by a range of a factor of 2 to a large factor of approximately 50 for some of the higher energy normal modes. Based on the good agreement between spectra plotted with LC- ω PBE and experiment, it appears that the long-range corrections to the XC potential are necessary for describing the electronic structure in systems where CT states are important.

Further evidence of the necessity of long-range corrections to the XC potential for the HMB/TCNE system are shown on the low-energy side of the resonance Raman spectra. For the normal mode at 165 cm^{-1} , it is found that the motion of the system involves vibrations where the HMB and TCNE molecule change relative distance from one another. In situations where molecules involved in a CT excitation change distance from one another, it has been shown that the excitation energy^{26,53} and the potential energy surface⁶⁰ are dependent on the distance R between the donor and acceptor molecule. As the intermolecular distance is changed for a CT state, it is expected that the interactions between the cationic donor molecule (HMB) and anionic acceptor molecule (TCNE) behave as $1/R$, so the potential energy surface should have that behavior. Resonance Raman spectra, and especially the excited-state gradients used for determining Δ_k^n values, are a useful probe of how well the excited-state potential energy surface is described by the XC functional.

For B3LYP, although the magnitude of the Δ_k^n for the intermolecular D–A stretch is comparable to that determined experimentally (2.80 from B3LYP compared with -3.80 from experiment), this is likely a result of fortuitous error cancellation because the rest of the Δ_k^n values are poorly described using this functional. The experimental result for the sign of Δ_k^n for this mode was rationalized based on the fitting procedure by including the nuclear coordinate dependence on the transition dipole moment (Herzberg–Teller terms).³⁷ Note that for a single contributing excited state, the sign of Δ_k^n is irrelevant as long as Duschinsky rotations are not accounted for,⁶¹ and also that experimentally measured resonance Raman spectra can only yield the

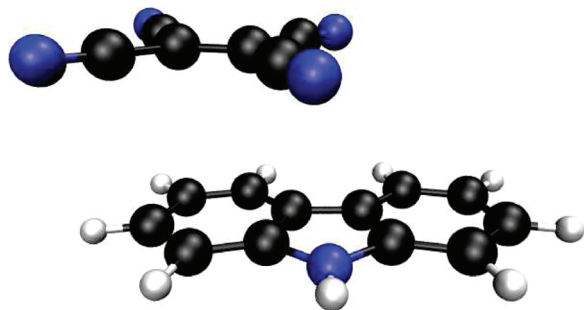


Figure 4. Optimized structure carbazole/TCNE complex using B3LYP/6-31G*.

Table 1. TDDFT Results for the Vertical Excitation Energies and Magnitudes of the Transition Dipole Moments Compared with the Experimental Estimates of These Values³⁷

method	CT1			CT2		
	λ (nm)	energy (eV)	$\ \mu^{0\eta}\ $ (au)	λ (nm)	energy (eV)	$\ \mu^{0\eta}\ $ (au)
B3LYP	1138	1.09	0.579	836	1.48	1.496
LC- ω PBE	549	2.26	0.740	487	2.55	0.785
experiment ³⁶	626	1.98	1.531	536	2.31	0.822

absolute value of Δ_k^n . The potential energy surface described by B3LYP displays a shape that is incorrect, likely attributed to the incorrect exponential decay of XC potential. LC- ω PBE has a slightly worse agreement for this mode with experiment (Δ_k^n is 2.12 for LC- ω PBE compared with -3.80 from experiment), but the rest of the spectrum is described much more similarly to the experimental results. This finding reflects the importance of asymptotic decay of the XC potential in this system.

It is also interesting to compare the methods based on the vibrational reorganization energy. The total reorganization energies for experiment,³⁷ B3LYP, and LC- ω PBE are 3517, 2140, and 2738 cm^{-1} , respectively. Comparing these values indicates both DFT methods underestimate the reorganization energies, by approximately 1400 and 800 cm^{-1} for B3LYP and LC- ω PBE, respectively. This partially reflects that both B3LYP and LC- ω PBE underestimate the value of Δ_k^n for almost every mode when compared with the fit determined experimentally. Another cause for the difference between

the Δ_k^n values is that those calculated from TDDFT are for the HMB/TCNE complex in vacuum, but those determined experimentally include the effects of solvent. Inclusion of solvent effects can have a dramatic effect on the resonance Raman spectrum for molecules involving CT states, as was shown in a previous study of rhodamine 6G.²⁵

Carbazole/TCNE. The optimized structure for the carbazole/TCNE complex is shown in Figure 4. Results for the calculation of vertical excitations and transition dipole moments using TDDFT with both XC functionals are compared with experimental findings in Table 1. For the experimental estimates for the locations of the two excitations, electronic structure calculations involving the Pariser–Parr–Pople (PPP)⁶² and semiempirical AM1 methods³⁶ were used to determine that the HOMO and HOMO-1 of carbazole are energetically similar. These analyses were performed for the carbazole donor in vacuum and, therefore, do not include interactions with the TCNE acceptor molecule. The calculations presented in this work include the entire complex as shown in Figure 4, and when TDDFT was applied using the LC- ω PBE functional to the carbazole/TCNE complex, it was determined that two CT states exist in energetically close proximity for this complex. When the calculation is performed using B3LYP the excitation energies are severely underestimated for both CT states, which has been observed previously in the literature and can be traced to self-interaction errors resulting from missing HF exchange.⁶⁰ These states will be referred to as the CT1 and CT2 states in the discussion that follows. Having two states in the absorbance band confirms what was proposed in refs 36 and 62.

Inspection of the transition dipole moments in Table 1 indicates that each method gives a different description of both excited states. This is reflected in the very different simulation of the absorbance spectrum (see Figure 5) for each method. For B3LYP, the CT2 state is more intense than the CT1 state, and as a result, the CT2 state must be positioned near the absorbance maximum to obtain a good description. The LC- ω PBE functional finds that the two states have similar intensity and are positioned nearly equidistant from the absorbance maximum. Experimentally, the transition dipole moments were determined as fitting parameters that gave the best agreement with the resonance Raman spectra.³⁶

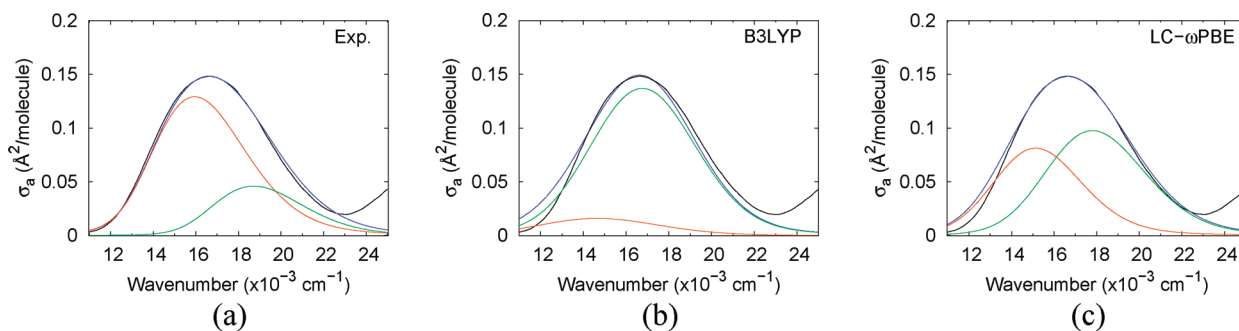


Figure 5. Absorbance spectra for carbazole/TCNE with fits compared to the experimental absorbance spectrum (black line, ref 36). In the figures, the total fit is the blue line, the fit for the CT1 state (at lower energy) is the red line, and the fit for the CT2 state (at higher energy) is the green line. The fit labeled “Exp.” used $\Gamma_{\text{CT1}} = 300 \text{ cm}^{-1}$, $\Gamma_{\text{CT2}} = 200 \text{ cm}^{-1}$, $\Theta = 1200 \text{ cm}^{-1}$, scale factor = 0.732. Using the B3LYP functional: $\Gamma_{\text{CT1}} = 250 \text{ cm}^{-1}$, $\Gamma_{\text{CT2}} = 300 \text{ cm}^{-1}$, $\Theta = 2200 \text{ cm}^{-1}$, scale factor = 0.893. For LC- ω PBE: $\Gamma_{\text{CT1}} = 300 \text{ cm}^{-1}$, $\Gamma_{\text{CT2}} = 250 \text{ cm}^{-1}$, $\Theta = 1600 \text{ cm}^{-1}$, scale factor = 2.01.

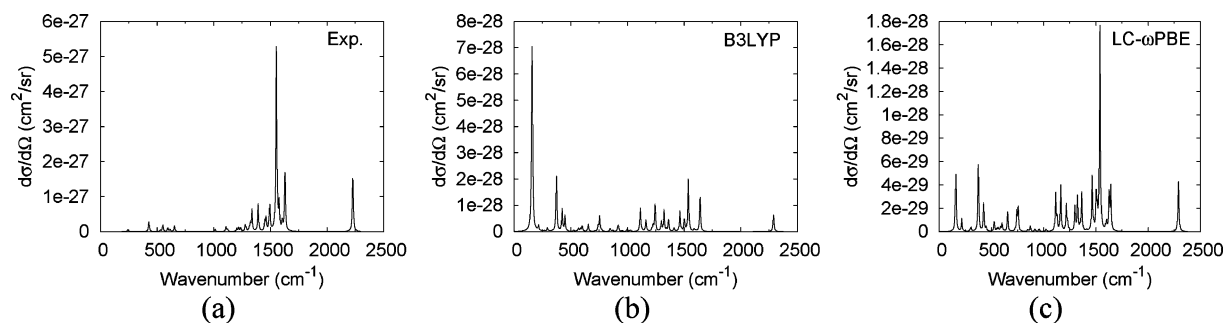


Figure 6. Resonance Raman spectra at 601 nm for carbazole/TCNE using the optimum modeling parameters for the absorbance spectra plotted in Figure 5. Data for the spectrum labeled “Exp.” are from ref 36. Spectra are broadened with a Lorentzian function with 10 cm^{-1} width.

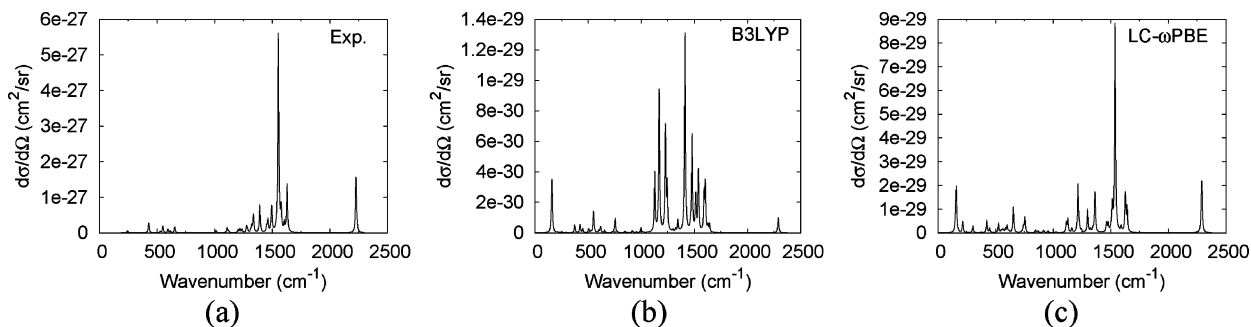


Figure 7. Individual resonance Raman spectra for the CT1 state at 601 nm for the carbazole/TCNE complex using the optimum modeling parameters for the absorbance spectra plotted in Figure 5. Data for the spectrum labeled “Exp.” are from ref 36. Spectra are broadened with a Lorentzian function with 10 cm^{-1} width.

In this case, it is found that the CT1 state is more intense and is positioned closer to the absorbance maximum than the CT2 state resulting from the transition dipole moments.

In Figure 5, plots of the fits using the experimental, B3LYP, and LC- ω PBE Δ_k^n values (see Tables 6–8 in the Supporting Information) are shown. For this system, the CT1 state involves the HOMO to LUMO excitation regardless of which functional is used. The CT2 state is a transition between the HOMO-1 and LUMO for both functionals. Both functionals yield similar character for the orbitals involved in both CT states, where the HOMO and HOMO-1 are localized on the carbazole molecule, and the LUMO is contributed to by the TCNE molecule. The simulated absorbance spectrum is especially good for the LC- ω PBE functional. However, both XC functionals yield a poorer fit than experimentally determined parameters for the red-edge of the absorbance maximum.

Simulated resonance Raman spectra at the absorbance maximum (601 nm) are shown in Figure 6. At that wavelength both CT states have some contribution to the total resonance Raman spectrum. A simple qualitative examination of the spectra indicates again that B3LYP overestimates the contribution of the intermolecular D–A stretch at 159.78 cm^{-1} . Even if that mode were ignored, the mode at 1537.44 cm^{-1} would not stand out on the resonance Raman spectrum even though that mode (predominantly the C=C stretch of the TCNE molecule) is the dominant feature observed experimentally. Clearly, the potential energy surfaces of the CT1 and CT2 states are poorly described by B3LYP, as indicated by the resonance Raman spectrum.

LC- ω PBE yields a spectrum where the correct mode is the dominant feature of the spectrum but also has some noticeable differences. The fact that the relative intensities of the modes at 1537.44 , 1602.16 (carbazole C–H and N–H in-plane bending and ring deformations), and 2287.67 cm^{-1} (TCNE CN stretch) are nearly identical to experiment demonstrates that this functional is giving a better description of the potential energy surfaces of the CT1 and CT2 states. Differences present on the spectrum below 1500 cm^{-1} between the LC- ω PBE functional and the experiment result in part from the energetic placement of the excited states to obtain good fits and also in particular the transition dipole moments for the two states. Because the experimental fit determines that CT2 has a significantly smaller transition dipole moment, this state will not be observed having as strong of an effect on the resonance Raman spectrum. LC- ω PBE finds that the transition dipole moments for both states have similar magnitude, causing a noticeable contribution from both states to the resonance Raman spectrum.

To determine the effects of the different excited states, resonance Raman spectra for the individual CT1 and CT2 (Figures 7 and 8) states were simulated by only including one excited state. Comparing the results for the CT1 state alone, it is clear that LC- ω PBE gives a very similar resonance Raman spectrum compared with the experimental result. It is also distinct that the description from B3LYP is poor, likely resulting from the same incorrect behavior of the XC potential for B3LYP described above for the HMB/TCNE complex. Only a few modes around 750 and 1250 cm^{-1} , involving carbazole ring deformations and in-plane bending

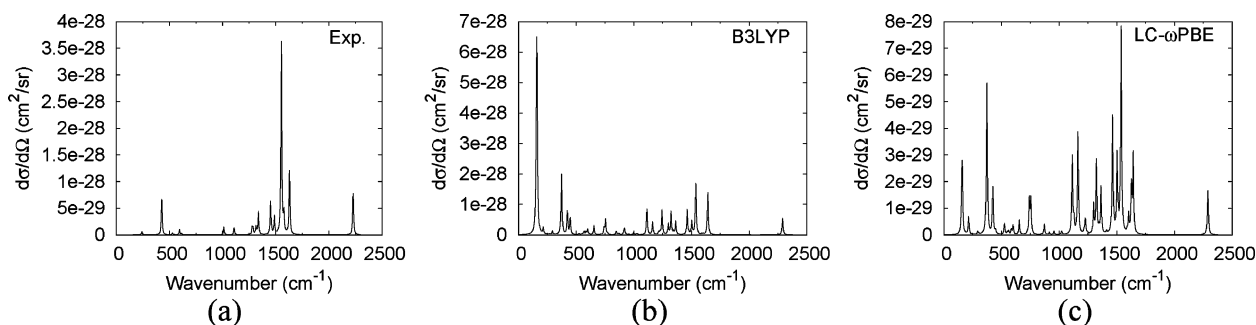


Figure 8. Individual resonance Raman spectra for the CT2 state at 601 nm for carbazole/TCNE complex using the optimum modeling parameters for the absorbance spectra plotted in Figure 5. Data for the spectrum labeled “Exp.” are from ref 36. Spectra are broadened with a Lorentzian function with 10 cm^{-1} width.

of C–H and N–H bonds, have intensities that are overestimated by the LC- ω PBE functional.

A comparison between the resonance Raman spectra for the individual CT1 and CT2 states from the LC- ω PBE functional and experimental data gives an indication of how well each excited state is described using the LC-DFT method. The CT1 state simulated by LC- ω PBE compares well with what is found from experiment, indicating that this state is described correctly by the LC- ω PBE functional. When the CT2 state is simulated, there is a clear difference between what is observed experimentally and using the LC- ω PBE functional, however, due to the appearance of two modes with high intensity near 425 cm^{-1} (carbazole out of plane ring deformation) and also several high-intensity modes between 1000 and 1500 cm^{-1} for the LC- ω PBE model. This disagreement stands out, but based on the agreement of LC- ω PBE with experiment for the CT1 state, it does not seem like the description of the CT2 state potential energy surface by the LC- ω PBE is necessarily incorrect.

Use of the experimental fit proposed in ref 36 would mean that observing both CT states separately is very difficult because both overlap on the blue edge of the absorbance maximum. Although interference effects³⁶ are observed between the two CT states, this does not seem to be the reason why the Δ_k^n values for the CT1 and CT2 states are so similar experimentally. Because of the large amount of overlap, the positioning of the two states and the large transition dipole moment of the CT1 state compared to the CT2 state from experiment, the CT1 state is observed to dominate the contributions to the total resonance Raman spectrum (Figure 6). As a result, any wavelength used to measure a resonance Raman spectrum for the carbazole/TCNE complex would have intensities largely derived from the CT1 state. Likely, the CT2 state was only partially observed due to changes in relative peak intensities on the total resonance Raman spectrum, but its contribution was so weak that it could not be fully resolved.

Because the CT1 and CT2 states overlap to some degree in the absorbance spectrum there is a possibility that interference effects may occur in the resonance Raman spectrum. Interference can be constructive and result in an increased total differential cross-section compared to the individual contributions of either CT states or destructive which causes a reduction in peak intensities for specific modes on the total resonance Raman spectrum.³⁹ These

effects are described in detail in ref 36 for parallel transition dipole moments and more generally in ref 39 for numerous situations involving different modeling parameters. Three important cases that may cause destructive interference are having: Δ_k^n values with opposite signs, the angle between the transition dipole moments of different states, and detuning from resonance.

The contribution from Δ_k^n values with opposite signs is important because for two contributing excited states the differential Raman scattering cross-section is proportional to the sum of the polarizability of each state squared. This causes terms in the differential Raman scattering cross-section that are products of the polarizability of each state that both depend on the Δ_k^n values as multiplicative factors. The angle between the transition dipole moments for the two states can result in components of the transition dipole moments that are opposite in sign. For the experimental spectrum, it was assumed that the transition dipoles are parallel,³⁶ but LC- ω PBE finds that there is a 125.5° angle between them. This means that only the first effect is present in the total resonance Raman spectrum for the experimental data, but both are present for the LC- ω PBE functional. The third effect was elaborated on in ref 39 and is caused by detuning from resonance with the excited states resulting in resonance deenhancement. This third effect can be observed when changing the excitation wavelength along the absorbance peak and is clearly viewed for the experimental data, where the sign of the Fourier integrals can change and result in different signs for the real and imaginary components of the polarizability. This effect is most clear when the sum-over-states expression for the Raman polarizability is examined, where having an excitation wavelength greater than the vertical excitation energy creates a negative sign in the energy denominator.

The interference effects, I , in terms of the resonance Raman cross-sections for this system can be quantified as

$$I = \left(\frac{d\sigma}{d\Omega}\right)_{\text{total}} - \left(\frac{d\sigma}{d\Omega}\right)_{\text{CT1}} - \left(\frac{d\sigma}{d\Omega}\right)_{\text{CT2}} \quad (13)$$

where the terms on the right-hand side are, respectively, the total differential Raman cross-section, the differential Raman cross-section for the CT1 state alone, and the differential Raman cross-section for the CT2 state alone. This allows plots of interference to be made over the spectral range of

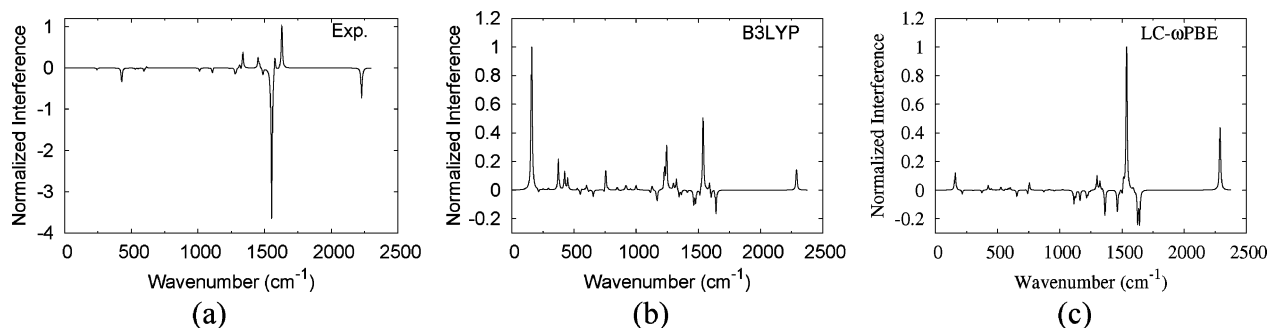


Figure 9. Plots of the interference between the CT1 and CT2 states at 601 nm using eq 13. Plots are normalized to the peak with maximum constructive interference. Data for the spectrum labeled “Exp.” are from ref 36.

the resonance Raman spectrum, like what is shown in Figure 9. For each plot, data are normalized to the mode with maximum constructive interference, and thus, any negative peak represents destructive interference.

An obvious difference is apparent by investigating the interference patterns presented in Figure 9. First, the figure derived from experimental data has constructive interference present for modes involving Δ_k^n values that are opposite in sign for the CT1 and CT2 states. This disagrees with the analysis presented in ref 36, where having Δ_k^n values with opposite signs for the CT1 and CT2 states was thought to be the only factor causing destructive interference. Because the transition dipole moments are parallel (directed along the positive z -axis), their components have the same sign and cannot result in destructive interference effects. This means that the third effect resulting from the sign of the integral must lead to the seemingly opposite picture for the constructive and destructive interference observed for the experimental data. A test calculation (data not shown) where the experimental absorbance maxima for CT1 and CT2 were shifted to the locations determined for the best fit of the LC- ω PBE functional caused the interference pattern to change from what is shown in Figure 9 to a pattern similar to that found for LC- ω PBE, indicating that this detuning effect is important for determining interference. Also, it can be shown that the interference pattern changes when data is compared at different excitation wavelengths for the experimental data (see Figure 1 in the Supporting Information).

For both the B3LYP and LC- ω PBE functionals, all three interference effects are present, resulting in a different description of the interference pattern. Although the magnitude of constructive interference for the B3LYP functional, particularly on the low-frequency side of the spectrum, does not agree with the description by LC- ω PBE, examination of the interference plots show that many of the same modes have constructive interference. This is surprising based on the poor description of the potential energy surfaces by B3LYP, but more importantly it demonstrates that all three effects discussed above are important for describing interference. It turns out that out of eight modes found experimentally to have destructive interference, six of these modes at 617.00, 1127.24, 1362.96, 1463.04, 1477.85, and 1626.19 cm^{-1} were found to show destructive interference using the LC- ω PBE functional.

One final detail is the vibrational reorganization energy for the carbazole/TCNE system. Experiment finds compa-

table values for the CT1 and CT2 states of 1988 and 1950 cm^{-1} , respectively. The respective values are 2441 and 1456 cm^{-1} for B3LYP and 1375 and 2606 cm^{-1} for LC- ω PBE. Due to the different descriptions of the two CT states using B3LYP and LC- ω PBE as compared with experiment, it is difficult to directly compare the individual reorganization energies. However, the sum of the reorganization energies for the two CT states are similar in all three cases.

Conclusions

This work presents the first application of the long-range corrected density functional theory (LC-DFT) method to resonance Raman scattering of donor–acceptor complexes. The popular global hybrid functional B3LYP and LC-functional LC- ω PBE were compared against experimental data for the carbazole/tetracyanoethylene (TCNE) and hexamethylbenzene/TCNE donor–acceptor complexes with important charge transfer (CT) states in their optical absorbance spectra. Using simulations of the absorbance and resonance Raman spectra involving the time-dependent formalism of Heller et al., it was found that, even though B3LYP could simulate absorbance spectra reasonably, it yielded poor descriptions of the excited-state potential energy surfaces in the FC region, as indicated by resonance Raman spectra. The LC- ω PBE functional simulates the absorbance spectra well and significantly improves the description of the potential energy surfaces in the FC region, as shown by its better agreement with the experimental resonance Raman spectrum for both complexes. For the carbazole/TCNE complex in particular the overlapping CT1 and CT2 states cause interference effects that change relative peak intensities on the resonance Raman spectrum. These effects can be traced to three factors: the sign of Δ_k^n for each state, the angle between the transition dipole moments, and detuning from resonance with each state. In the analysis we showed that all three factors need to be accounted for. Finally, the total vibrational reorganization energy from both B3LYP and LC- ω PBE was compared to what was calculated experimentally using the Marcus theory of electron transfer.⁹ Both functionals yield reasonable predictions of the total vibrational reorganization energy, but LC- ω PBE distributes single-mode contributions similarly to what was found experimentally based on the resonance Raman spectra. In order to improve agreement with the experimental data, it might be necessary to include solvent effects in these systems so that the Δ_k^n

values compare more closely with experiment for the CT states investigated in this work.²⁵

Acknowledgment. L.J. would like to thank Dr. Niri Govind for many interesting discussions regarding the long-range functionals. L.J. acknowledges the CAREER program of the National Science Foundation (Grant No. CHE-0955689) for financial support, start-up funds from the Pennsylvania State University (Penn State), and support received from Research Computing and Cyberinfrastructure, a unit of Information Technology Services at Penn State. This research was supported in part by the National Science Foundation through TeraGrid resources provided by NCSA under grant number (TG-CHE090144).

Supporting Information Available: Data tables containing the Δ_k^H values, single-mode reorganization energies, normal mode assignments from the experimental and B3LYP data, and figures showing the interference patterns at several excitation wavelengths. This information is available free of charge via the Internet at <http://pubs.acs.org/>.

References

- Wang, X.; Maeda, K.; Chen, X.; Takanabe, K.; Domen, K.; Hou, Y.; Fu, X.; Antonietti, M. *J. Am. Chem. Soc.* **2009**, *131*, 1680.
- Wang, X.; Maeda, K.; Thomas, A.; Takanabe, K.; Xin, G.; Carlsson, J. M.; Domen, K.; Antonietti, M. *Nat. Mater.* **2009**, *8*, 76.
- Brabec, C. J.; Zerzaa, G.; Cerullo, G.; Silvestri, S. D.; Luzzatic, S.; Hummelend, J. C.; Sariciftci, S. *Chem. Phys. Lett.* **2001**, *340*, 232.
- Page, C. C.; Moser, C. C.; Chen, X.; Dutton, P. L. *Nature* **1999**, *402*, 47.
- Liu, Z.; Zhang, C.; He, W.; Qian, F.; Yang, X.; Gao, X.; Guo, Z. *New J. Chem.* **2010**, *34*, 656.
- Tseng, T.-C.; et al. *Nature Chem.* **2010**, *2*, 374.
- Hupp, J. T.; Williams, R. D. *Acc. Chem. Res.* **2001**, *34*, 808.
- Barbara, P. F.; Meyer, T. J.; Ratner, M. A. *J. Phys. Chem.* **1996**, *100*, 13148.
- Marcus, R. A. *J. Chem. Phys.* **1965**, *43*, 679.
- Myers, A. B. *Chem. Rev.* **1996**, *96*, 911.
- Biswas, N.; Umaphathy, S. *Chem. Phys. Lett.* **1998**, *294*, 181.
- Hoekstra, R. M.; Zink, J. I.; Telo, J. P.; Nelsen, S. F. *J. Phys. Org. Chem.* **2009**, *22*, 522.
- Albrecht, A. C. *J. Chem. Phys.* **1961**, *34*, 1476.
- Albrecht, A. C.; Hutley, M. C. *J. Chem. Phys.* **1971**, *55*, 4438.
- Tang, J.; Albrecht, A. C. *J. Chem. Phys.* **1968**, *49*, 1144.
- Tannor, D. J.; Heller, E. J. *J. Chem. Phys.* **1982**, *77*, 202.
- Heller, E. J. *Acc. Chem. Res.* **1981**, *14*, 368.
- Heller, E. J.; Sundberg, R.; Tannor, D. *J. Phys. Chem.* **1982**, *86*, 1822.
- Lee, S.-Y.; Heller, E. J. *J. Chem. Phys.* **1979**, *71*, 4777.
- Petrenko, T.; Neese, F. *J. Chem. Phys.* **2007**, *127*, 164319.
- Neese, F.; Petrenko, T.; Ganyushin, D.; Olbrich, G. *Coord. Chem. Rev.* **2007**, *251*, 288.
- Kelley, A. M. *J. Phys. Chem. A* **2008**, *112*, 11975.
- Kane, K. A.; Jensen, L. *J. Phys. Chem. C* **2010**, *114*, 5540.
- Guthmuller, J.; Champagne, B. *J. Chem. Phys.* **2007**, *127*, 164507.
- Guthmuller, J.; Champagne, B. *J. Phys. Chem. A* **2008**, *112*, 3215.
- Dreuw, A.; Weisman, J. L.; Head-Gordon, M. *J. Chem. Phys.* **2003**, *119*, 2943.
- Bernasconi, L.; Sprik, M.; Hutter, J. *J. Chem. Phys.* **2003**, *119*, 12417.
- Neugebauer, J.; Louwse, M. J.; Baerends, E. J.; Wesolowski, T. A. *J. Chem. Phys.* **2005**, *122*, 094115.
- Lange, A.; Herbert, J. M. *J. Chem. Theory Comput.* **2007**, *3*, 1680.
- Iikura, H.; Tsuneda, T.; Yanai, T.; Hirao, K. *J. Chem. Phys.* **2001**, *115*, 3540.
- Vydrov, O. A.; Scuseria, G. E. *J. Chem. Phys.* **2006**, *125*, 234109.
- Yanai, T.; Tew, D. P.; Handy, N. C. *Chem. Phys. Lett.* **2004**, *393*, 51.
- Livshits, E.; Baer, R. *Phys. Chem. Chem. Phys.* **2007**, *9*, 2932.
- Baer, R.; Neuhauser, D. *Phys. Rev. Lett.* **2005**, *94*, 043002.
- Stein, T.; Kronik, L.; Baer, R. *J. Am. Chem. Soc.* **2009**, *131*, 2818.
- Egolf, D. S.; Waterland, M. R.; Kelley, A. M. *J. Phys. Chem. B* **2000**, *104*, 10727.
- Markel, F.; Ferris, N. S.; Gould, I. R.; Myers, A. B. *J. Am. Chem. Soc.* **1992**, *114*, 6208.
- Kulinowski, K.; Gould, I. R.; Myers, A. B. *J. Phys. Chem.* **1995**, *99*, 9017.
- Shin, K.-S. K.; Zink, J. I. *J. Am. Chem. Soc.* **1990**, *112*, 7148.
- Wootton, J. L.; Zink, J. I. *J. Am. Chem. Soc.* **1997**, *119*, 1895.
- Kramers, H. A.; Heisenberg, W. *Z. Phys.* **1925**, *31*, 681.
- Dirac, P. A. M. *Proc. R. Soc. London, Ser. A* **1927**, *114*, 710.
- Neugebauer, J.; Reiher, M.; Kind, C.; Hess, B. A. *J. Comput. Chem.* **2002**, *23*, 895.
- Craig, D. P.; Thirunamachandran, T. *Molecular Quantum Electrodynamics: An Introduction to Radiation Molecule Interactions*; Dover Publications, Inc.: Mineola, NY, 1998; pp128–141.
- Long, D. A. *The Raman Effect: A Unified Treatment of the Theory of Raman Scattering by Molecules*; John Wiley & Sons, Ltd.: West Sussex, England, 2002; pp 85–152, 221–270.
- Neugebauer, J.; Hess, B. A. *J. Chem. Phys.* **2004**, *120*, 11564.
- Reiher, M.; Neugebauer, J.; Hess, B. A. *Z. Phys. Chem.* **2003**, *217*, 91.
- Straatsma, T.P.; Apra, E.; Windus, T.L.; Bylaska, E.J.; de Jong, W.; Hirata, S.; Valiev, M.; Hackler, M.; Pollack, L.; Harrison, R.; Dupuis, M.; Smith, D.M.A.; Nieplocha, J.; Tipparaju V.; Krishnan, M.; Auer, A.A.; Brown, E.; Cisneros, G.; Fann, G.; 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.; Jae, J.; Johnson, B.; Ju, J.; Kobayashi, R.; Kutteh, R.; Lin, Z.; Littlefield, R.; Long, X.; Meng, B.; Nakajima, T.; Niu, S.; Rosing, M.; Sandrone, G.; Stave, M.; Taylor, H.; Thomas, G.; van Lenthe, J.; Wong, A.; Zhang, Z. *NWChem, A Computational Chemistry Package for Parallel Computers*, version 5.1, a modified version; Pacific Northwest National Laboratory: Richland, WA, 2007.
- (49) Jensen, L.; Govind, N. *J. Phys. Chem. A* **2009**, *113*, 9761.
- (50) Govind, N.; Valiev, M.; Jensen, L.; Kowalski, K. *J. Phys. Chem. A* **2009**, *113*, 6041.
- (51) Henderson, T. M.; Janesko, B. G.; Scuseria, G. E. *J. Chem. Phys.* **2008**, *128*, 194105.
- (52) Weintraub, E.; Henderson, T. M.; Scuseria, G. E. *J. Chem. Theory Comput.* **2009**, *5*, 754.
- (53) Rohrdanz, M. A.; Martins, K. M.; Herbert, J. M. *J. Chem. Phys.* **2009**, *130*, 054112.
- (54) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648.
- (55) Bauernschmitt, R.; Ahlrichs, R. *Chem. Phys. Lett.* **1996**, *256*, 454.
- (56) Bauernschmitt, R.; Häser, M.; Treutler, O.; Ahlrichs, R. *Chem. Phys. Lett.* **1996**, *264*, 573.
- (57) Rohrdanz, M. A.; Herbert, J. M. *J. Chem. Phys.* **2008**, *129*, 034107.
- (58) Grimme, S. *J. Comput. Chem.* **2004**, *25*, 1463.
- (59) Grimme, S. *J. Comput. Chem.* **2006**, *27*, 1787.
- (60) Dreuw, A.; Head-Gordon, M. *J. Am. Chem. Soc.* **2003**, *126*, 4007.
- (61) Myers, A. B.; Pranata, K. S. *J. Phys. Chem.* **1989**, *93*, 5079.
- (62) Landman, U.; Ledwith, A.; Marsh, D. G.; Williams, D. J. *Macromolecules* **1976**, *9*, 833.

CT1002779

Real-Time TD-DFT Simulations in Dye Sensitized Solar Cells: The Electronic Absorption Spectrum of Alizarin Supported on TiO₂ Nanoclusters

Rocío Sánchez-de-Armas, Jaime Oviedo López, Miguel A. San-Miguel, and Javier Fdez. Sanz*

Department Physical Chemistry, University of Seville (Spain)

Pablo Ordejón and Miguel Pruneda

CIN2—Centre d'Investigació en Nanociència i Nanotecnologia (CSIC-ICN), Campus UAB, 08193 Bellaterra, Spain

Received June 2, 2010

Abstract: The structural and electronic properties of the alizarin dye supported on TiO₂ nanoclusters have been examined by means of time-dependent density-functional (TD-DFT) calculations performed in the time-domain framework. The calculated electronic absorption spectrum of free alizarin shows a first band centered at 2.67 eV that upon adsorption features a red shift by 0.31 eV, in agreement with both experimental and previous theoretical work. This red shift arises from a relative stabilization of the dye LUMO when adsorbed. To analyze the dependence of the electronic properties of the dye-support couple on the size of metal-oxide nanoparticles, different models of (TiO₂)_n nanoclusters have been used (with $n = 1, 2, 3, 6, 9, 15,$ and 38). As a conclusion, the minimal model is good enough to theoretically reproduce the main feature in the spectrum (i.e., the energy shift of the main band upon binding to TiO₂). However, it fails in creating intermediate states which could play a significant role under real experimental conditions (dynamics of the electronic transfer). Indeed, as the size of the nanocluster grows, the dye LUMO moves from the edge to well inside the conduction band (Ti 3d band). On the other hand, to assess the consistency of the time-domain approach in the case of such systems, conventional (frequency-domain) TD-DFT calculations have been carried out. It is found that, as far as the functional and basis set are equivalent, both approaches lead to similar results. While for small systems the standard TD-DFT is better suited, for medium to large sized systems, the real-time TD-DFT becomes competitive and more efficient.

1. Introduction

Dye sensitized solar cells (DSSC) have attracted considerable attention in recent years.¹ They are based on transition metal derivatives or organic dye molecules that are adsorbed on a semiconductor, generally a metal oxide such as TiO₂. The photochemical properties of different sensitizers have been extensively investigated, in an attempt to design dyes with maximal visible light absorption coupled to long-lived

excited states. Ruthenium dyes are the most efficient sensitizers in DSSCs by now, although they have some disadvantages. Molecular modification of these dyes is an arduous task due to their complicated synthesis routes, and the materials involved in the synthesis procedures are rather expensive. In contrast, organic dyes have a much lower cost, and their molecular design is more convenient and easier. They have large absorption coefficients due to intramolecular $\pi \rightarrow \pi^*$ transitions, and there are no concerns about limited resources, because they do not contain noble metals such as Ru.

* Corresponding author e-mail: sanz@us.es.

A key process in the operation of this kind of devices is the charge injection from the dye molecule adsorbed at the nanoparticle surface to the conduction band states of the semiconductor. The injection rate has been found to depend on the electronic properties of both the dye and the semiconductor, as well as the distance between them.² The DSSCs can be classified into two types, depending on the electron injection mechanism from the dye to the semiconductor. In type-I cells, the injection mechanism is indirect and first involves the photoexcitation to an excited state of the dye, from which an electron is transferred to the conduction band of the semiconductor. In contrast, type-II cells involve a direct mechanism or a “one-step” electron injection from the ground state of the dye to the conduction band of the semiconductor. A new charge-transfer band that, upon adsorption, appears in the electronic absorption spectrum of the dye reveals the direct injection mechanism, whereas in the case of an indirect mechanism, no new bands are seen in the spectrum.^{1,3}

Alizarin is an organic molecule that has been investigated in great detail as a photosensitizer, with high incident photon-to-current conversion efficiencies.^{1–7} The electron injection from the alizarin excited state into TiO₂ proceeds at an ultrafast 6 fs time scale.^{8,9} The experimental electronic absorption spectrum for free alizarin shows a low energy band centered at 2.88 eV (431 nm) and other, stronger bands at energies higher than 5 eV (250 nm) with a shoulder at 3.82 eV (325 nm). Upon adsorption, the lowest band is red-shifted by about 0.4 eV and appears at 2.47 eV (503 nm), which is characteristic of an indirect mechanism. The other bands are also shifted and appear at energies higher than 3.55 eV (350 nm).⁴

An important issue of this process is the nature of the excited states, in particular, whether the photoexcitation results in a state mainly localized at the chromophore, i.e., a locally excited state, or a state that is mainly localized at the titanium, i.e., a charge-transfer state. Another issue under discussion is whether the electron transfer (ET) mechanism is adiabatic or nonadiabatic.^{1,2,6,10} In the adiabatic mechanism, the coupling between the dye and the semiconductor is large, and ET occurs through a transition state along the reaction coordinate that involves a concerted motion of nuclei. During an adiabatic transfer, the electron formally remains in the same Born–Oppenheimer (adiabatic) state that continuously changes its localization from the dye to the semiconductor along the reaction coordinate. Nonadiabatic effects decrease the amount of ET that happens at the transition state but open up a new channel involving direct transitions from the dye into the semiconductor that can occur at any nuclear configuration. In this case, a strong coupling is unnecessary as a large density of TiO₂ acceptor states can still give rise to an ultrafast injection even when the coupling is weak. The nonadiabatic ET effect is a quantum effect and, similar to tunnelling, shows exponential dependence on the donor–acceptor separation.

Theoretical methods are a powerful tool for molecular design, and conclusions drawn from calculations are valuable guidelines for the synthesis of new efficient dyes. Time-dependent density functional theory (TD-DFT) has been

usually chosen to reproduce the electronic absorption spectra of organic dyes. This method provides an improved treatment of the electron correlation effects relative to other methods like CIS.⁴ Nevertheless, calculations with conventional linear response (LR) TD-DFT implementation in the frequency domain¹¹ to estimate individual excitations are computationally demanding, and the results are limited to relatively small models. The alizarin molecule is one of the smallest sensitizer dyes, and that makes it a good system to be theoretically investigated.^{12,13}

In previous work by Duncan and Prezhdo,⁴ the free and adsorbed alizarin spectra were calculated using TD-DFT. A minimal model was used in which the dye molecule was bound to a single Ti atom and a set of water and hydroxyl ligands. An essentially localized excitation of the dye followed by an electron transfer to a localized surface state and a spreading into the conduction band states were suggested. Results from this and similar work¹² are able to qualitatively show the main features of experimental spectra, and the effect of bounding alizarin to the surface, but they are limited in size and make it difficult to fully assess the role of electronic delocalization into the clusters. In addition, extensive theoretical studies on the charge transfer dynamics have been performed.^{1,6,13} From these works emerged the conclusion that the adiabatic mechanism was more dominant relative to the nonadiabatic one. Recent experimental work,⁷ however, concluded that finite size effects produce a multiple electron injection with different time scales, which is characteristic of a nonadiabatic event.

In this work, we report on density functional theory (DFT) calculations on the alizarin molecule isolated and adsorbed on TiO₂ clusters. First, we consider the geometry of the organic dye in the electronic ground state as well as the geometries of the dye-cluster models. Second, we explore the optical response of the different structures using TD-DFT. These calculations are performed in a complementary way through the conventional linear response approach¹⁴ and through real time propagation^{15,16} with a basis of localized orbitals. One important initial goal of this work is to establish the accuracy of the real time TD-DFT method within the SIESTA code in the case of organic dyes. For this reason, we have chosen a molecule that has been extensively studied. Since the cost of this method is relatively low, we are able to extend the study to moderately large systems that are closer to experimental conditions. The aim in this case is to analyze the electronic structure as the cluster size increases and how this affects to the theoretical spectra. We will focus on the delocalization of the excited electrons into the clusters and how it might play a significant role in the process. We conclude that the real-time TD-DFT calculated absorption spectra are in good agreement with the experimental results, in terms of both absolute absorption energies and relative band intensities. Moreover, it is shown that, although a minimal model can account for the qualitative features of the spectra, it might be too limited in describing the electronic structure of the system, which is essential in describing the dynamics of the electron transfer.

2. Computational Methods

2.1. Models. The system to be studied involves an organic molecule that is adsorbed on TiO₂ nanoparticles suspended in an electrolyte. It is often implicitly assumed that each experimental absorption peak corresponds to a single adsorption geometry. However this does not necessarily have to be the case. For instance, there could be differences in the cluster size, adsorption sites, alizarin isomers, etc., so the actual spectra would be the average of an enormous ensemble of states.³ Here, we present a set of models that try to cover the main aspects of this issue.

The majority of the experimental studies on the interfacial ET in the dye sensitized TiO₂ systems are carried out with TiO₂ nanoparticles that are a mixture of the rutile and anatase crystal forms of TiO₂ with a variety of exposed surfaces. However, many experimental studies use colloidal TiO₂ particles that are prepared by the hydrolysis of TiCl₄ in cold water. The anatase phase becomes more stable than the rutile phase when the sizes of TiO₂ clusters are smaller than 14 nm.¹⁷

It is theoretically unfeasible to determine the actual adsorption geometry for the alizarin adsorbed on TiO₂ clusters in solution, as it would be necessary first to know the cluster morphology. We have made calculations adsorbing the alizarin molecule on clusters of several sizes. It is well beyond the scope of this paper to find the absolute energy minimum for the clusters, even more so when the minima could change upon adsorption and/or including solvation effects. The starting geometries for the TiO₂ clusters in this study were taken from the literature. The small models^{18,19} are taken from optimized geometries starting from spherically shaped particles, whereas the largest one (U₃₈) started from the anatase structure.²⁰ We regard these clusters as a reasonable choice for their actual geometries.

The selected models range from a minimal cluster that was used in previous theoretical work to a relatively large one with 38 TiO₂ units. The minimum model (as in ref 4) corresponds to a neutral TiO₄H₄ cluster, which includes a single metal atom. Formally, it can be written as U₁ (2H₂O), where U is a TiO₂ unit. In addition, other clusters labeled as U₂, U₃, U₆, U₉, U₁₅, and U₃₈ were also considered. The larger cluster has a diameter of about 1 nm that has to be compared to nanoparticles of about 2–6 nm used in experiments, and therefore this model will provide results more comparable to real experiments. However, the use of small clusters is interesting in the study of the effect of delocalization and also helps to establish which is the minimal size to obtain reliable data.

It has been suggested from the infrared data and theoretical structure calculations that catechol, whose structure is similar to that of alizarin, is covalently linked to the surface Ti atom in a chelating bidentate mode involving the oxygen atoms of its two hydroxyl groups. Following Duncan and Prezhdo,⁴ the present study uses the bidentate geometry for the adsorbed alizarin as a first working model (Figure 1).

Another crucial aspect is whether the alizarin (or the whole system) is neutral or charged upon adsorption. We have studied two extreme cases for the adsorption: the system was

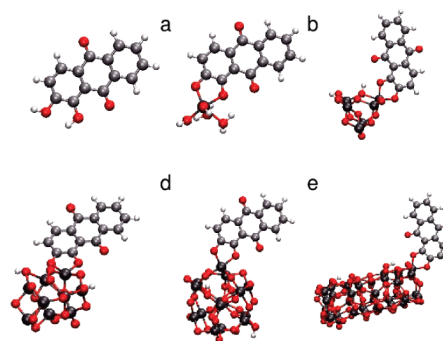


Figure 1. Optimized geometries for free alizarin (a) and adsorbed alizarin on several models (b–f): minimal model U₁, U₆, U₉, U₁₅, and U₃₈, respectively.

either neutral (denoted as Aliz-2H-Cluster), and therefore the system keeps the two phenolic hydrogen atoms, or dianionic (Aliz-Cluster-2), where these hydrogen atoms were removed. For the neutral case, the hydrogen atoms are linked to cluster oxygen atoms rather than to the alizarin molecule, as preliminary calculations showed that this might be the most stable case, although we did not try every possible adsorption site rearrangement. As shown in Figure 1, in the models we have used, the hydrogen atoms were as separated as possible on the clusters, following the strategy used in ref 13. Again, we should stress that this is only a sensible model for the adsorption. Generally speaking, the neutral models are not appropriate for small clusters, since adsorbing hydrogen atoms on them will significantly deform the geometry. On the other hand, we found that for the charged models the calculations are increasingly difficult to converge as the size grows. We should take into account that in real working conditions, the charge is counter-balanced by the electrolyte ions and solvent molecules, which are not explicitly included in our calculations. In some specific cases, we have made an estimation of solvation effects through a polarization continuum model (PCM). All the structures were fully optimized prior to calculating the electronic absorption spectra.

2.2. Electronic Absorption Spectra: TD-DFT in Real Time. The electronic absorption spectra were simulated from TD-DFT calculations through two sets of complementary calculations. Although it is known that TD-DFT can significantly underestimate the energies of long-range charge transfer states, that is not the case with the present calculation, in which the photoexcited state shows a moderate and relatively short-ranged charge transfer.² The Gaussian 03 code²¹ was used to perform conventional LR-TD-DFT calculations¹¹ and the SIESTA^{22,23} code to perform real time TD-DFT calculations.^{15,16} Since the latter method has been less used, we include a detailed description in this subsection.

In the simplest case, the lowest band observed in an experimental electronic absorption spectrum comes from a HOMO to LUMO excitation. In the traditional LR-TD-DFT method, a key input parameter is the number of excitations to be included in the calculation. When the excitation that is responsible for the main band in the spectrum is the HOMO–LUMO or mixed with excitations from an occupied orbital close to the HOMO to a virtual orbital close to the LUMO (i.e: HOMO–1 to LUMO, HOMO to LUMO+1,

etc.), then only a small number of excitations have to be included in the calculation. However, when there are molecular orbitals between the initial and the final state of the electronic excitation (i.e: HOMO to LUMO+15), then a large number of excitations will be energetically possible at a lower energy than responsible for the main band. These transitions have usually negligible intensities but have to be explicitly included in the calculations (by allowing a larger number of excitations). As a rule of thumb, if the excitation is from the HOMO orbital to the LUMO+ N , the number of excitations to be included is N^2 . In the free alizarin molecule, the transition is basically HOMO to LUMO; however, when the molecule is adsorbed, there are many states in between the origin and the destiny of the excitation because of the presence of the cluster. As a consequence, performing a conventional (frequency domain) LR-TD-DFT for large systems becomes prohibitive at a certain size (see the Results section for a case with definite numbers).^{24,25} In contrast, the “real time” TD-DFT method generates all the possible excitations at the same time (by applying an electric field to the molecule). The spectrum is calculated from a molecular dynamic simulation and extends to very high excitations (although the spectrum has no real meaning above the ionization energy). The method is not computationally cheap for small systems, as it requires relatively long simulation times, but becomes competitive against the traditional one as the size of the system grows. One disadvantage of this approach is that we cannot fully characterize the nature of the different transitions because the implementation of the method is based on electron density, and there is no information about the excited states wave functions.

The method we have employed¹⁶ involves the description of the electronic states using a linear combination of atomic orbitals (LCAO). Because the size of the LCAO basis is small, the TD-DFT calculations can be done efficiently using the techniques described below. The use of the LCAO basis leads to matrices with a size considerably smaller than when other basis sets are used or when real-space grid methods are employed. The scheme is based on the SIESTA code,²² which is used to compute the initial wave functions and the Hamiltonian matrix for each time step. Core electrons are replaced by norm-conserving pseudopotentials²⁶ in the fully nonlocal Kleinman–Bylander²⁷ form, and the basis set is a general and flexible linear combination of numerical atomic orbitals (NAOs), constructed from the eigenstates of the atomic pseudopotentials.^{28–30} The NAOs are confined, being strictly zero beyond certain radii. In addition, the electron wave functions and density are projected onto a real-space grid in order to calculate a certain contribution to the Hamiltonian matrix, such as the Hartree and exchange-correlation terms (details are given in ref 22).

We carry out the calculations in the time domain, explicitly evolving the wave functions, following the approach of ref 16. We consider as an initial state a bound system in a finite electric field; i.e., the Hamiltonian includes a perturbation $\Delta H = -E \cdot x$. For the linear response calculations in the present work, we have set the value of this field to 0.1 V/Å. The system is solved for the ground state using standard time independent DFT. Then, we switch off the electric field at

time $t = 0$, and for every subsequent time step we propagate the occupied Kohn–Sham eigenstates by solving the time-dependent Kohn–Sham equation

$$i\hbar \frac{\partial \psi}{\partial t} = H\psi \quad (1)$$

where H is the time-dependent Hamiltonian given by

$$H = -\frac{\hbar^2}{2m}\nabla^2 + V_{\text{ext}}(r, t) + e^2 \int \frac{\rho(r', t)}{|r-r'|} dr' + V_{\text{xc}}[\rho](r, t) \quad (2)$$

with $V_{\text{ext}}(r, t)$ being the external (ionic) potential, $\rho(r, t)$ being the electron density, and $V_{\text{xc}}[\rho](r, t)$ being the exchange-correlation potential. The calculation of the exchange-correlation potential is done using the general gradient approximation (GGA). For every time step, we propagate the wave function using a Crank–Nicholson scheme,¹⁶ and from the new wave functions we construct the new density matrix. The electron density is then obtained and used for the calculation of the Hamiltonian in the new cycle and to calculate the dipole moment $D(t)$. After a Fourier transformation, we obtain the elements of the frequency-dependent polarizability tensor $\alpha_{ij}(\omega)$, which defines the first order response of the system: $D(\omega) = \alpha(\omega) E(\omega)$. Using a step function external electric field $E(t) = E \Theta(-t)$, we have

$$\text{Im}[\alpha(\omega)] = \omega \frac{\text{Re}[D(\omega)]}{E} \quad (3)$$

The polarizability determines the optical properties and in particular can be used to calculate the dipole strength function: $S(\omega)$

$$S(\omega) = \frac{2m}{\pi e^2 \hbar} \omega \text{Im}[\alpha(\omega)] \quad (4)$$

The dipole strength function, $S(\omega)$, is proportional to the photoabsorption cross-section and allows for a direct comparison with experiments.

2.3. Technical Details. The Gaussian 03 calculations were performed using two different functionals, B3LYP and PBE, and the 6-31G** basis set. To analyze the solvent effect, we have done the calculations both in the gas phase and containing the methanol solvation effect using the polarizable continuum model (PCM). When necessary, up to the 100 lowest singlet transitions were included in the calculations. The SIESTA calculations were performed using the PBE generalized gradient approach functional together with Troullier–Martins pseudopotentials and an auxiliary real-space grid equivalent to a plane-wave cutoff of 130 Ry. A nonstandard DZP basis set of NAOs has been used.³¹ A time step of 1.5×10^{-3} fs was used for the time evolution, and the simulation was carried out for a total time of 60 fs, allowing for an appropriate energy conservation.

3. Results

We start this section with the simulated electronic absorption spectrum of neutral free alizarin (model Aliz-2H). The computed spectrum from real-time TD-DFT calculations is reported in Figure 2a. As can be seen, the lowest band

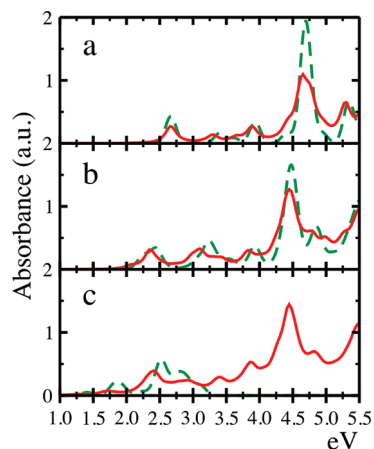


Figure 2. Comparison between conventional (frequency domain) and real-time TD-DFT (red full line) theoretical spectra. From top to bottom: the free alizarin and the adsorbed alizarin on the U_1 and U_6 models.

Table 1. Lowest Band Energies (eV) for Free and Adsorbed Alizarin from Standard (Frequency Domain) TD DFT Calculations Using the PBE and B3LYP Functionals, with and without Including the Solvent Effect through a PCM Model^a

model	PBE	B3LYP	PBE (PCM)	B3LYP (PCM)
Aliz-2H	2.67 (-0.21)	3.25 (0.37)	2.44 (-0.44)	2.99 (0.11)
Aliz-2H- U_1 (2H ₂ O)	2.26 (-0.21)	2.96 (0.49)	2.19 (-0.28)	2.63 (0.16)
Aliz-2H- U_6	2.48 (0.01)	2.99 (0.52)		

^a Experimental bands for free and adsorbed alizarin are at 2.88 eV and ~2.47 eV, respectively (differences from the experimental data in parentheses).

appears at 2.67 eV, in agreement with the experimental measurement at 2.88 eV. This band is followed by a series of low intensity features and ends with a strong absorption at about 4.65 eV. In order to compare this result with that obtained from LR-TD-DFT calculations, we also report in Figure 2a the spectrum estimated using the same exchange-correlation functional, PBE (see also Table 1). As can be observed, the similarity is remarkable despite the differences in the approaches (basis sets and all-electrons versus pseudo-potentials).³²

Bonding the dye to a titania cluster induces a red-shift of the lowest band, as observed in the two series of spectra reported in Figure 2b and c that correspond to the minimal model Aliz-2H- U_1 (2H₂O) and a medium sized cluster, Aliz-2H- U_6 , respectively. The estimated red shift is 0.31 and 0.27 eV, in good agreement with the experiment (0.41 eV). With the aim of extending the above comparison to these models, these plots also include the spectra obtained from standard TD-DFT approach. In the case of the minimal model (Figure 2b), the agreement also is notorious, while for model Aliz-2H- U_6 some differences are clearly observed, mainly in the upper energy region of the spectrum (above 3.5 eV). It is worth noting that the conventional spectra extend only to a small energy range because of the limited number of transitions included in the calculation. For instance, the number of excitations to reach the energy of the main band grows from 3 in the smallest model to 56 in

the Aliz-2H- U_6 model. From our calculations, we estimate that, to reach the main band energy, more than 300 excitations should be included in the U_{15} model and more than 1000 in the U_{38} case. This is due to the large number of virtual orbitals located on the cluster that produces electronic transitions with negligible intensities (see Table 2 and Figure 5). Since in our calculations we have limited the maximum number of excitations to 100, the upper energy region of the spectrum is inadequately described, and, in fact, that is why the spectrum suddenly fades at 3.35 eV in the case of model Aliz-2H- U_6 .

An additional point to bear in mind in the analysis and comparison with the experimental data concerns the solvent effects as well as the choice of the exchange-correlation functional. The functional choice is the key parameter when calculating the electronic spectra (Table 1). The position of the main peak moves to higher energies when passing from the PBE to the B3LYP functional. The energy increments are 0.58, 0.70, and 0.49 eV for the three models here considered. Compared to the experimental spectra, the PBE functional underestimates the transition energy while the values obtained with the hybrid B3LYP functional appear overestimated. The inclusion of solvent effects through a PCM model tends to reduce further the excitation energies (about 0.2–0.3 eV). Therefore, the PBE frequencies are in worse agreement and the B3LYP frequencies in better agreement when including solvent effects.

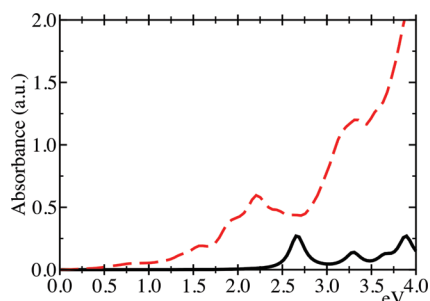
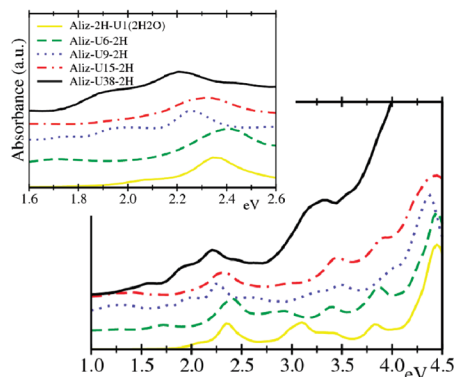
The overall behavior of the real time TD-DFT spectra for all the models we have considered is summarized in Figures 3 and 4. In general, the simulated spectra look similar, and as can be seen, upon adsorption there is a significant red shift of the main band of alizarin which moves from 2.67 eV for the free molecule to 2.21 eV when adsorbed on the U_{38} cluster (−0.46 eV theoretical shift compared to the −0.41 experimental shift). There is also an increase in the absorbance in the visible region because of the high number of low energy excitations within the cluster. The spectrum of our benchmark model, Aliz- U_{38} -2H, reported in detail in Figure 3, shows broad bands centered at 2.21, 3.27, and 4.2 eV. These bands also appear for other models, although we note that there are small changes in the position and relative intensities. In general, the smaller the model, the more featured the spectra. As the system grows bigger, the bands in the spectra become wider and less resolved because the number of involved excitations increases. This is in agreement with what is found in experiments and also reflects the fact that the model system is switching from having discrete energy levels to energy bands. The main feature of the spectra (the red shift upon adsorption) is already reproduced, albeit qualitatively, by the minimal model. However, to account for the delocalization of the excitations over the titania particle, the detailed analysis of the electronic transitions shows that larger cluster models are needed.

In Figure 5, the computed Kohn–Sham orbital energy levels for several models in the relevant energy window are shown. The HOMO orbital energy has been taken as a reference to allow a better comparison and has been set to zero. In the free alizarin, the most intense contribution to the lowest energy band comes from an excitation from the

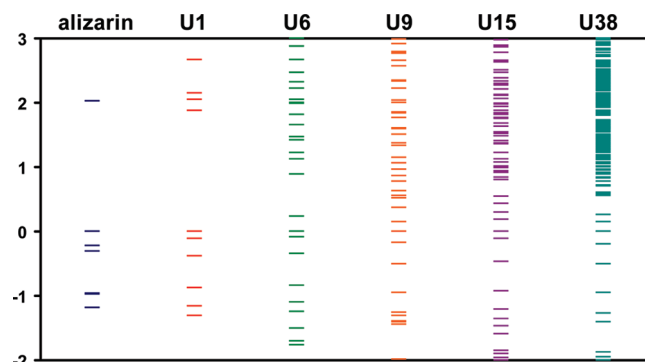
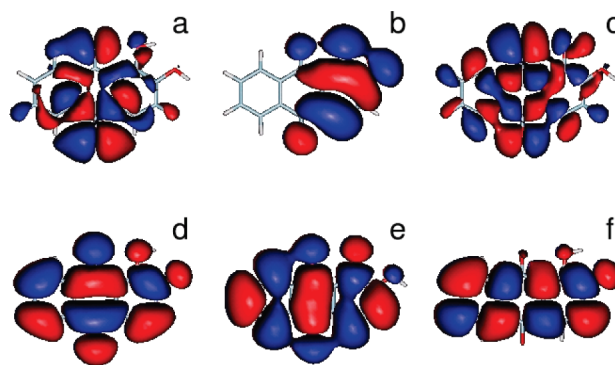
Table 2. Assignments of Electronic Excitations for Free Alizarin and Adsorbed on U₁ and U₆ Models^a

transition energy (eV), excitation number	oscillator strength	wave function
Free Alizarin		
2.67 (3)	0.0856	H-1→L (0.63), H-1→L+3 (-0.13), H-3→L+2 (0.11)
U ₁ Model		
2.18 (2)	0.0098	H→L+1 (0.58), H→L (0.31), H-1→L (-0.19)
2.26 (3)	0.0282	H→L+2 (0.46), H→L (-0.32), H-1→L+1 (-0.24), H→L+1 (0.20), H-1→L (0.16), H-1→L+2 (-0.15)
2.29 (4)	0.0105	H-1→L+1 (0.62), H→L+2 (0.19), H→L (-0.17), H-1→L+2 (0.16)
U ₆ Model		
2.48 (56)	0.0368	H→L+10 (0.39), H-3→L+6 (-0.30), H→L+11 (0.20), H→L+13 (-0.19), H-2→L+11 (0.18), H-2→L+12 (0.17), H→L+14 (-0.12)
2.53 (58)	0.0147	H-5→L+4 (0.58), H-3→L+6 (-0.27), H-1→L+14 (0.18), H-2→L+10 (0.11), H→L+10 (-0.11)
2.53 (59)	0.0177	H-5→L+5 (0.49), H-5→L+4 (-0.30), H-3→L+6 (-0.30), H-1→L+14 (0.17), H→L+10 (-0.12)
2.53 (60)	0.0213	H-5→L+3 (0.50), H-3→L+6 (0.32), H-5→L+4 (0.24), H-1→L+14 (-0.19), H→L+10 (0.13), H-2→L+12 (-0.11)
2.56 (63)	0.0214	H-1→L+14 (0.60), H-3→L+6 (0.17), H→L+10 (0.16), H→L+14 (0.14), H-2→L+10 (-0.13)

^a Only selected transitions with enough oscillator strength around the main peak are included (H stands for HOMO and L for LUMO).

**Figure 3.** Real-time TD-DFT theoretical spectra for free alizarin (black full line) and adsorbed on U₃₈.**Figure 4.** Real-time TD-DFT theoretical spectra for the adsorbed alizarin on several models. Bottom to top: U₁, U₆, U₉, U₁₅, and U₃₈. The spectra have been shifted along the y axis to allow for a better comparison. Inset: a zoom in of the main band energy region.

HOMO-1 to the LUMO (Table 2). This is in agreement with previous theoretical calculations making use of the PW91 functional.⁴ The HOMO-1 orbital is a π orbital localized over the hydroxyl part of the molecule (Figure 6). The lone electron pairs of the hydroxyl and quinone oxygens contribute to this orbital. The LUMO is distributed over the whole molecule ring. The difference in the dipole moments between the ground and excited states was experimentally

**Figure 5.** Molecular orbital energies for the ground state. From left to right: the free alizarin and the adsorbed U₁, U₆, U₉, U₁₅, and U₃₈ models. For the sake of comparison, the levels have been shifted to set the HOMO energy as the zero energy.**Figure 6.** Selected frontier occupied and virtual molecular orbitals of free alizarin. (a) HOMO-2, (b) HOMO-1, (c) HOMO, (d) LUMO, (e) LUMO+1, (f) LUMO+2.

quantified as 4.4 D, which reflects the significant intramolecular transfer of electronic charge in the electronic transition.³

After binding to titanium, there is a redistribution of electron density. In the minimal model, the relevant excitations that form the main band start from the HOMO or

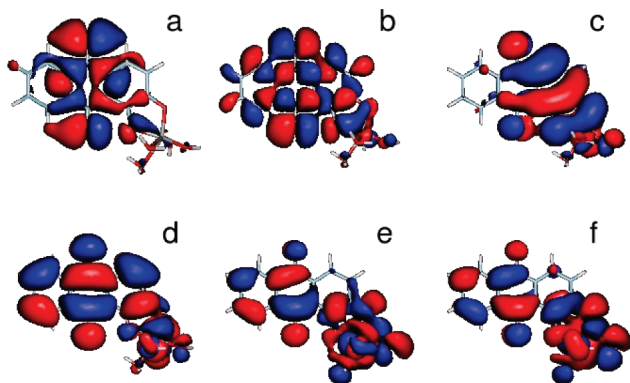


Figure 7. Selected frontier occupied and virtual molecular orbitals of the minimal model. (a) HOMO–2, (b) HOMO–1, (c) HOMO, (d) LUMO, (e) LUMO+1, (f) LUMO+2.

HOMO–1 to the LUMO, LUMO+1, and LUMO+2 since the orbital energies of these occupied and virtual states respectively are close in energy (Table 2 and Figure 5). Since the electronic states shown in Figure 7 extend over both the molecule and the Ti, there are more excitations producing a nonzero contribution to the main band. The HOMO and LUMO resemble those of the free molecule (HOMO–1 and HOMO switch positions when passing to the minimal model). However, in every case, there is a noticeable contribution from the Ti atom. This is especially true in the LUMO+1 and LUMO+2 where the d orbitals of Ti play a significant role. The occupied electronic states are mainly localized on the alizarin molecule whereas the virtual electronic states show a larger contribution from the Ti 3d orbitals; that is, the molecular orbitals extend over the whole system. It has been suggested from Stark effect measurements that the excited state of alizarin would be essentially localized on the dye molecule, and the nanoparticle electric field would cause the absorption band shift.³ This does not seem to agree with our calculations, since we conclude that the shift in the spectrum of the minimal model seems to be related to a stabilization of the acceptor states relative to the donor states because of the hybridization.

When the cluster contains several titanium atoms, the electronic wave function may spread over the whole system so becoming more stable than the orbital localized on the alizarin. In Figure 8, some relevant molecular orbitals for the Aliz-2H–U₆ model are shown. These orbitals have been chosen because they are the most important in contributing to the lowest energy band. In this case (Table 2), the band is made of several optical transitions, and each of them corresponds to several pairs of donor and acceptor orbitals. The occupied orbitals that contribute most are HOMO–5, HOMO–3, HOMO–1, and HOMO. They all are mainly localized on the alizarin, although there is some minor contribution from the Ti atoms close to the molecule (i.e., in the HOMO). On the other hand, LUMO+10 and LUMO+11 are the main contributors to the most intense transition and are localized on the alizarin. They resemble the LUMO in the free molecule. LUMO+6, which shows an important contribution from the alizarin, is responsible for a small band at about 1.9 eV (Figure 2). From Table 2 it is clear that LUMO, LUMO+1, LUMO+2, LUMO+7,

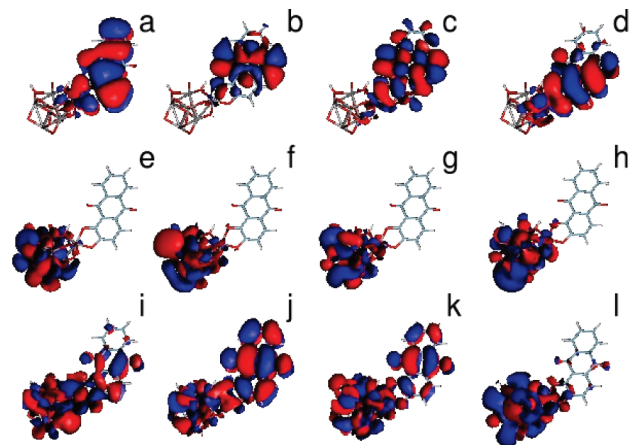


Figure 8. Selected frontier occupied and virtual molecular orbitals of the U₆ model. (a) HOMO–3, (b) HOMO–2, (c) HOMO–1, (d) HOMO, (e) LUMO, (f) LUMO+1, (g) LUMO+2, (h) LUMO+3, (i) LUMO+6, (j) LUMO+10, (k) LUMO+11, and (l) LUMO+12.

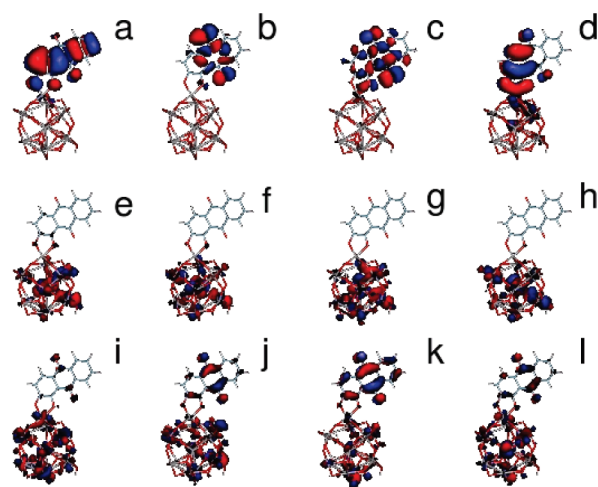


Figure 9. Selected frontier occupied and virtual molecular orbitals of the U₁₅ model. (a) HOMO–3, (b) HOMO–2, (c) HOMO–1, (d) HOMO, (e) LUMO, (f) LUMO+1, (g) LUMO+2, (h) LUMO+3, (i) LUMO+30, (j) LUMO+32, (k) LUMO+33, and (l) LUMO+35.

LUMO+8, and LUMO+9 do not participate in the transitions that make up the main band because they are totally localized on the cluster, so their overlap with the occupied orbitals is negligible.

On the other hand, when the cluster size becomes larger, the virtual orbitals close to the HOMO are well localized at the center of the cluster (see LUMO and LUMO+1 in Figure 9). The energies of these orbitals show some “molecular” character since they are discrete. Above these states, there are manifold states delocalized on the whole cluster that become closer to a band (a continuum of states) for the U₃₈ model (Figure 5). It is worth notice that there is an increasing number of virtual orbitals between the HOMO and LUMO of the free alizarin as the system grows bigger. The former LUMO transforms into LUMO+10 for the U₆, LUMO+17 for the U₉, LUMO+33 for the U₁₅, and about LUMO+90 for the U₃₈. In every case, the energy difference remains about 2 eV, which makes the main band in the spectrum

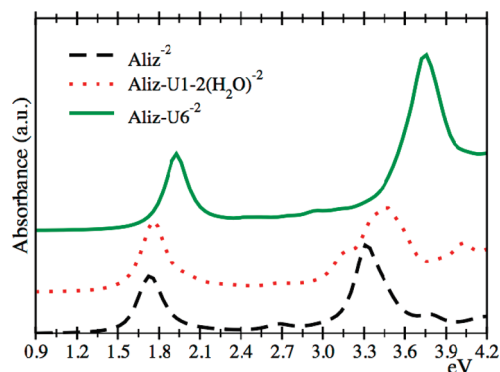


Figure 10. Real-time TD-DFT theoretical spectra for isolated dianion alizarin (black dashed line) and adsorbed on U_1 (red dot line) and U_6 (green full line).

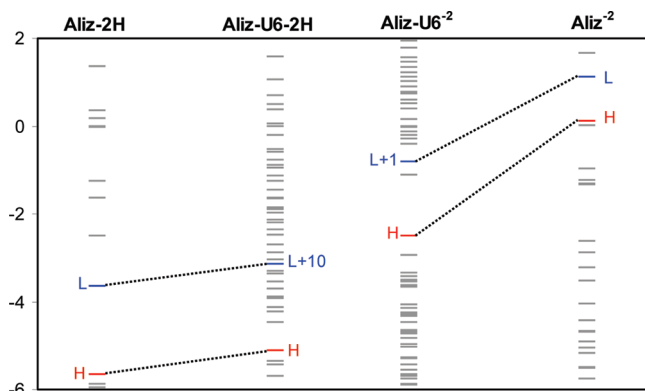


Figure 11. Molecular orbital energies of neutral and dianion alizarin forms isolated and adsorbed on the U_6 model.

almost unaltered in position (Figure 4). The number of excitations below the main transition grows substantially, although the intensities are usually close to zero.

An additional point we have considered concerns the fact that the lowest absorption in the alizarin electronic spectrum moves when the pH of the solution changes. Thus, in an ethylene glycol/water solution the position of this peak is 2.85, 2.34, and 2.17 eV at pH of 3, 10, and 14 respectively.³ This shift is interpreted in terms of the actual form of alizarin that can be neutral, monoanionic and dianionic depending on the pH. To analyze this point additional calculations for negatively charged models were performed for both isolated and supported alizarin species. For isolated alizarin, the lowest peak for the dianion form is computed at 1.75 eV and moves to 1.87 eV for the monoanion species (0.8 eV below the neutral), in agreement with the experimental trend. When the dianion species is supported on a titania cluster, a blue shift of this band is observed as shown in Figure 10, where the spectra of the dianion form isolated and bound to U_1 and U_6 models are depicted. This behavior agrees with the experimental shift observed in the spectra of the TiO_2 -alizarin complex at pH = 11.³ The electronic changes involved in these shifts might be understood taking into account the redistribution of the energy levels that happens upon dissociation and adsorption. As shown in Figure 11, double proton dissociation leads to a relative destabilization of HOMO with respect to LUMO (both strongly destabilized). As a consequence, the first band that, as already

commented upon, is mainly contributed by this excitation is red-shifted. Upon adsorption, both HOMOs and LUMOs are stabilized, although the lowering of the former is larger. Consequently, the band is blue-shifted. Notice that the stabilization of the alizarin molecular orbitals that occurs upon adsorption is consistent with a larger redistribution of the negative charge over the titania cluster. Such redistribution is more efficient when the size of the cluster increases, and therefore the blue shift for U_6 is larger than for U_1 . Although larger clusters should lead to larger shifts, we could not go further due to convergence problems. Finally, to end this discussion, it is worth noting that although the results are in qualitative agreement with experiments (in the sense that there is a blue shift of the main peak upon adsorption),³ we should stress again that the system does not include solvent or cationic entities that could counterbalance the net charge on the system.

A final point worth highlighting is that recent experimental work on the electron injection process revealed multiexponential dynamics with different time constants, characteristic of a nonadiabatic process.⁷ There were electron transfer events in about 0.1 ps and back electron transfer in 0.2 ps. In addition, a second injection time in about 17 ps and a slow back electron transfer in 1 ns were reported. This observation was explained on the basis of a finite size effect, which could lead to discreteness in the conduction band levels leading to different injection times to different levels within the conduction band.⁷ From Figure 5, it is clear how the energy level bands are formed as the system grows while maintaining some discrete levels above the HOMO localized on the cluster.

We speculate that there are electron injections to discrete levels localized on the cluster and to levels corresponding to LUMO in the free molecule. The direct and back electron transfer are fast for the former case, since the levels are localized and the energy difference between donor and acceptor levels is small. On the other hand, in the latter case, the direct electron transfer is slightly slower (the energy difference is larger), whereas the back electron transfer is very slow since the electron wave function is spread over the whole cluster, which favors the transfer of the electron to TiO_2 . This second electron transfer is responsible for the solar cell response. The possibility of changing the energy levels by using certain cluster sizes could have a potential use in the design of more efficient organic solar cells.

Conclusions

In this study, we have analyzed the electronic structure and the electronic absorption spectra of neutral alizarin both free and bound to TiO_2 nanoclusters of different sizes. To this aim, we have employed two different implementations of the TD-DFT to simulate the electronic spectra: the time domain and the frequency domain approaches. A first conclusion that can be reached is that, within the range of models we have examined, both methods provide comparable results as far as the basis set and the exchange-correlation functional are similar. This proves that the real-time TD-DFT scheme implemented within the SIESTA code is accurate for this type of system.

For free alizarin, our simulated spectra show a low energy band at 2.67 eV, in agreement with the experimental value (2.88 eV) and previous theoretical work. Upon adsorption, and also in agreement with the experiment, this band features a shift toward lower energies of 0.27–0.46 eV (PBE) depending on the model size. The effect of the size of the TiO₂ nanoparticle has been modeled by using differently sized nanoclusters going from the minimal TiO₄H₄ unit to a (TiO₂)₃₈ cluster, for which the estimated red shift is found to be 0.46 eV, in excellent agreement with the experiment (0.41 eV). From the analysis, we conclude that the minimal model is enough to reproduce the main feature in the spectra (a red-shift upon binding). Larger clusters do not much modify this picture since the energy difference between the HOMO and the orbital corresponding to the free molecule LUMO remains almost the same (i.e., the main excitation stays almost unchanged). However, the electronic structure of the system does change for larger systems. The virtual orbital on the alizarin responsible for the excitation moves from being below the “conduction band” for the minimal model to being well into the conduction band for larger clusters. In addition, we note that there are finite size effects that create discrete levels localized on the clusters. This may have important consequences for the dynamics of electron transfer. From our study, we regard the a unit of (TiO₂)₆ as the smallest nanocluster model able to simulate semiquantitatively all the features in the electronic structure of the system.

Acknowledgment. This work was funded by the Spanish Ministerio de Ciencia e Innovación, MICINN, projects MAT2008-4918, CSD2008-0023, FIS2009-12721-C04-01, and CSD2007-00050. R.S.A. thanks the Junta de Andalucía for a predoctoral grant (P08-FQM-3661 and EXC/2005/FQM-1126). Part of the calculations have been carried out at the Barcelona Supercomputing Center—Centro Nacional de Supercomputación (Spain).

References

- Duncan, W. R.; Prezhdo, O. V. *Annu. Rev. Phys. Chem.* **2007**, *58*, 143.
- Duncan, W. R.; Stier, W. M.; Prezhdo, O. V. *J. Am. Chem. Soc.* **2005**, *127*, 7941.
- Nawrocka, A.; Krawczyk, S. *J. Phys. Chem. C* **2008**, *112*, 10233.
- Duncan, W. R.; Prezhdo, O. V. *J. Phys. Chem. B* **2005**, *109*, 365.
- Duncan, W. R.; Craig, C. F.; Prezhdo, O. V. *J. Am. Chem. Soc.* **2007**, *129*, 9–8528.
- Duncan, W. R.; Prezhdo, O. V. *J. Am. Chem. Soc.* **2008**, *130*, 9756.
- Kaniyankandy, S.; Verma, S.; Mondal, J. A.; Palit, D. K.; Ghosh, H. N. *J. Phys. Chem. C* **2009**, *113*, 3593.
- Huber, R.; Moser, J. E.; Grätzel, M.; Wachtveitl, J. *J. Phys. Chem. B* **2002**, *106*, 6494.
- Huber, R.; Spörlein, S.; Moser, J. E.; Grätzel, M.; Wachtveitl, J. *J. Phys. Chem. B* **2000**, *104*, 8995.
- Duncan, W. R.; Prezhdo, O. V. *J. Phys. Chem. B* **2005**, *109*, 17988.
- Casida, M. E. Recent Developments and Applications of Modern Density Functional Theory. In *Theoretical and Computational Chemistry*; Seminario, M., Ed.; Elsevier: Amsterdam, 1996; Vol 4, p 391.
- Kondov, I.; Wang, H.; Thoss, M. *Int. J. Quantum Chem.* **2006**, *106*, 1291.
- Guo, Z.; Liang, W. Z.; Zhao, Y.; Chen, G. H. *J. Phys. Chem. C* **2008**, *112*, 16655.
- Stratmann, R. E.; Scuseria, G. E.; Frisch, M. J. *J. Chem. Phys.* **1998**, *109*, 8218.
- Yabana, K.; Bertsch, G. F. *Phys. Rev. B* **1996**, *54*, 4484.
- Tsolakidis, A.; Sánchez-Portal, D.; Martin, R. M. *Phys. Rev. B* **2002**, *66*, 235416.
- Zhang, H.; Banfield, J. F. *J. Mater. Chem.* **1998**, *8*, 2073.
- Qu, Z.; Kroes, G. *J. Phys. Chem. B* **2006**, *110*, 8998.
- Hamad, S.; Catlow, C. R. A.; Woodley, S. M.; Lago, S.; Mejías, J. A. *J. Phys. Chem. B* **2005**, *109*, 15741.
- Lundquist, M. J.; Nilsing, M.; Persson, P.; Lunell, S. *Int. J. Quantum Chem.* **2006**, *106*, 3214.
- Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Rob, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03*; Gaussian, Inc.: Wallingford, CT, 2003.
- Soler, J. M.; Artacho, E.; Gale, J. D.; García, A.; Junquera, J.; Ordejón, P.; Sánchez-Portal, D. *J. Phys. Condens. Matter.* **2002**, *14*, 2745.
- Sánchez-Portal, D.; Ordejón, P.; Artacho, E.; Soler, J. M. *Int. J. Quantum Chem.* **1997**, *67*, 453.
- Different schemes have been proposed, however, to calculate optical spectra in the frequency domain, that rely on iterative methods to estimate the full spectrum of large systems with moderate computational workload (see ref 25). These are still not broadly used and have not been considered in our study.
- Walker, B. A. M.; Saitta, A. M.; Gebauer, R.; Baroni, S. *Phys. Rev. Lett.* **2006**, *96*, 113001.
- Troullier, N.; Martins, J. L. *Phys. Rev. B* **1991**, *43*, 1993.
- Kleinman, L.; Bylander, D. M. *Phys. Rev. Lett.* **1982**, *48*, 1425.
- Artacho, E.; Sánchez-Portal, D.; Ordejón, P.; García, A.; Soler, J. M. *Int. J. Quantum Chem. Status Solidi B* **1999**, *215*, 809.
- Junquera, J.; Paz, O.; Sanchez-Portal, D.; Artacho, E. *Phys. Rev. B* **2001**, *64*, 235111.

- (30) Anglada, E.; Soler, J. M.; Junquera, J.; Artacho, E. *Phys. Rev. B* **2002**, *66*, 205101.
- (31) The Ti atom basis set consists of 15 basis functions: two radial functions to represent the 4s electrons with cutoff radii 6.10 and 5.12 au, one radial function for the 4p orbitals with radius 3.11 au, and two functions for the 3d shell with radii 5.95 and 4.75 au. Oxygen electrons were described with 13 functions: two for the 2s shell with confinement radii 4.46 and 2.51 au, two functions for the 2p orbital with radii 6.17 and 2.33 au, and one function for 3d polarization shell with radius 5.06 au. The C atom basis set consists of 13 basis functions: two for the 2s states with radii 5.52 and 3.10 au,

two for the 2p orbital with radii 6.91 and 3.03 au, and one function for the 3d polarization shell with radius 5.12 au. Finally, for the H atom, five functions were used: two radial functions for the 1s shell with radii 4.95 and 1.77 au and one function for the 2p polarization shell with a 5.07 au radius.

- (32) The standard TD-DFT (Gaussian 03 code) spectrum was fitted by assigning a Gaussian function to each electronic excitation frequency with an area proportional to the oscillator strength of that excitation. A single width parameter was chosen so the band broadening was similar to that in real time TD-DFT spectra.

CT100289T

How Can Hydrophobic Association Be Enthalpy Driven?

Piotr Setny,^{*,†,‡,¶} Riccardo Baron,^{*,†,¶} and J. Andrew McCammon[†]

Department of Chemistry and Biochemistry, Center for Theoretical Biological Physics, Howard Hughes Medical Institute, Department of Pharmacology, University of California, San Diego and Physics Department, Technical University Munich, 85748 Garching, Germany

Received June 5, 2010

Abstract: Hydrophobic association is often recognized as being driven by favorable entropic contributions. Here, using explicit solvent molecular dynamics simulations we investigate binding in a model hydrophobic receptor–ligand system which appears, instead, to be driven by enthalpy and opposed by entropy. We use the temperature dependence of the potential of mean force to analyze the thermodynamic contributions along the association coordinate. Relating such contributions to the ongoing changes in system hydration allows us to demonstrate that the overall binding thermodynamics is determined by the expulsion of disorganized water from the receptor cavity. Our model study sheds light on the solvent-induced driving forces for receptor–ligand association of general, transferable relevance for biological systems with poorly hydrated binding sites.

1. Introduction

It is becoming widely recognized that water—the common environment for most biological processes—plays a significant role in binding (thermo)dynamics, being far more important than just a passive, embedding medium. This is particularly relevant for hydrophobic interactions, in which water-related effects are typically regarded as the origin of the entropic driving force. Nevertheless, our knowledge about how water structure and dynamics evolve during such binding events and what are the underlying thermodynamic contributions is still sparse.

One of the recent, key advances is the distinction among different length scales involved.^{1–4} It is now generally accepted that in the limit of small solutes, e.g., small hydrocarbons, the surrounding water structure and hydrogen bonding are not considerably affected. In this case, hydration thermodynamics appears to be dominated by the entropic effect of restricting spontaneous solvent fluctuations to only those permitting the presence of the solute.^{5–8} The resulting

solvent-mediated interactions, exemplified by the well-investigated case of methane pair in water,^{9–12} display a characteristic, entropy-stabilized free energy minimum for the direct contact of two solutes. However, the spontaneous assembly of small isolated hydrophobic molecules is not observed,¹³ owing to a comparatively larger configurational volume of the solvent-separated pairs.

Hydrophobic-driven association usually involves at least one interacting partner carrying an extended nonpolar region. It affects the hydrogen-bonding network of interfacial water molecules, inducing a variety of large-length scale hydrophobic effects.^{14,15} The actual solvent behavior critically depends on the strength of solute–solvent attraction and varies from the persistent hydration of strongly interacting objects,^{16,17} through the formation of a thin vapor-like interface next to flat or moderately concave hydrophobic surfaces,^{18–20} to the complete dewetting of sterically hydratable regions.^{21,22} The importance of such effects for the assembly of nanoscopic bodies was investigated in a number of molecular dynamics (MD) simulation studies.^{23–34} Attempts to characterize the underlying thermodynamic signatures in this context were, however, limited to relatively simple solutes, like plates or ellipsoids,^{23,28,33} and support the conventional view of entropy-driven hydrophobic association.

* Corresponding authors. E-mail: piotr.setny@tum.de (P.S.), rbaron@mccammon.ucsd.edu (R.B.).

[†] University of California, San Diego.

[‡] Technical University Munich.

[¶] These authors contributed equally to this work.

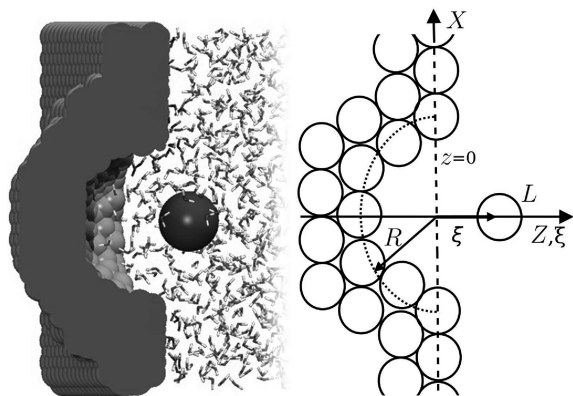


Figure 1. Snapshot and schematic representation of the explicitly solvated hemispherical cavity and spherical ligand (L) used in this study. Note that ($\xi = 0$) corresponds to the wall surface.

Is this conventional picture directly transferable to biologically relevant binding scenarios? They typically involve a receptor cavity, whose degree of hydration may vary significantly depending on geometry and composition,¹⁵ and a ligand, whose hydration belongs to the aforementioned small-length scale regime. Following an elegant reasoning by Dunitz,³⁵ it is usually assumed that the release of receptor-bound water is entropically favorable. This interpretation may hold even in the case of hydrophobic cavities, in which water is weakly bound. In such case, MD simulations indicated the formation of hydrogen-bonded water clusters,^{21,36–38} whose displacement upon binding should lead to the expected entropy dominated hydrophobic association. On the other hand, fundamental insight from nuclear magnetic resonance experiments suggests that water molecules permeating protein cavities may be also disordered^{39,40} with their displacement accompanied instead by an entropy loss,⁴⁰ in line with a recent MD study.⁴¹ These important findings remain largely unexplored in the context of cavity–ligand association, possibly contributing to unexpected effects, like enthalpy-driven binding of nonpolar ligands to the highly hydrophobic pocket of mouse major urinary protein (MUP).^{42–44}

Here, we investigate the driving forces for model nonpolar cavity–ligand association (Figure 1).^{30,45} This type of hydrophobic binding is peculiar in that it simultaneously involves both small- and large-scale alternative hydration regimes. Using explicit solvent MD, we derive the potential of mean force (PMF) for cavity–ligand interaction at five different temperatures. From the temperature dependence of the free energy, we obtain a complete picture of thermodynamic signature profiles along the association coordinate, thus far limited to simple solutes. As both associating partners are devoid of internal degrees of freedom, we directly capture water thermodynamic contributions and relate them to the ongoing changes in solute hydration. To our knowledge this is the first study elucidating the interplay between enthalpic and entropic components in the context of model cavity–ligand binding. We propose an explanation on why the observed association appears enthalpy driven, contrarily to what is usually expected for different hydrophobic systems.

2. Methods

2.1. Molecular Model and MD Simulations. The hemispherical cavity of 0.8 nm radius was embedded in a rectangular paraffin-like wall (Figure 1), constructed as a hexagonal close packed (HCP) grid (lattice constant 0.125 nm) of pseudoatoms interacting through 6–12 Lennard-Jones (LJ) potential with parameters $\epsilon_p = 0.0024 \text{ kJ mol}^{-1}$ and $\sigma_p = 0.4152 \text{ nm}$. Note that despite a small ϵ_p value, the well depth for flat wall–water interaction potential is 2.93 kJ mol^{-1} due to tight packing of wall particles, and hence, the wall should be regarded as realistic hydrophobic material (see the Supporting Information for details). The ligand was modeled as one neutral LJ sphere (methane parameters: $\epsilon_m = 1.23 \text{ kJ mol}^{-1}$ and $\sigma_m = 0.373 \text{ nm}$), and the TIP4P model was used for water molecules.⁴⁷ Pairwise LJ interactions were treated using standard mixing rules. The system used in MD simulations consisted of two identical $3.5 \times 3.3 \text{ nm}^2$ pocketed walls with adjacent ligands, mirrored along the Z axis such that they made two opposite sides of a 3 nm thick box filled with 1030 water molecules. Periodic boundary conditions were enforced with a box size equal to the system size in X,Y directions and 10 nm size in Z direction, i.e., with vacuum behind the walls. Association of each ligand with its pocket was sampled independently along ξ (the system symmetry axis) from 1.1 nm (in the bulk region) to -0.4 nm (inside the pocket) over 31 consecutive windows (0.05 nm apart) using the umbrella sampling method.⁴⁸ The corresponding PMF, $W(\xi)$, was calculated using the weighted histogram analysis method.⁴⁹ In all simulations, the distance between the two ligands was at least 2 nm (i.e., when one of them was in the bulk region, the other occupied the pocket interior), and it was assured that their movements were not correlated. All simulations were carried out with the CHARMM software.⁵⁰ For additional details we refer to refs 30 and 45. Five sets of independent simulations were performed at $T = 298, 308, 318, 328, \text{ and } 338 \text{ K}$. Simulation for each umbrella potential window was preceded by a number of 1.1 ns equilibrating runs during which the wall separation was iteratively adjusted until water density at the center of the solvent box matched the experimental value at the given temperature and pressure of $P = 1 \text{ bar}$ ⁵¹ with tolerance $\pm 2 \text{ g/L}$. The actual production runs (1 ns, preceded by 100 ps of equilibration) corresponded to *NVT* conditions with individual system volumes for each sampling window, in order to avoid barostat artifacts due to the presence of constrained walls (see ref 52 on this point). Owing to the system symmetry, effectively 2 ns sampling was obtained per ξ value, per temperature. Water density distribution maps were generated using xfarbe.⁵³

2.2. Free Energy, Entropy, and Enthalpy Changes and Their Uncertainties. The Gibbs free energy for the system with the ligand at a given ξ value reads

$$G(\xi) = W(\xi) + G(\infty) \quad (1)$$

where $G(\infty)$ was set to 0 by shifting to 0 the average G for $\xi \in [0.95, 1.1] \text{ nm}$. The corresponding entropy, $S(\xi)$, can be obtained through the partial temperature derivative of the free energy. The corresponding enthalpy reads $H(\xi) = G(\xi) + TS(\xi)$.

The water contribution to the total Gibbs free energy was obtained as $G_w(\xi) = G(\xi) - U_{CL}(\xi)$, where U_{CL} is the direct ligand–cavity interaction. Ligand–water, U_{LW} , and cavity–water, U_{CW} , interaction energies were calculated using force field terms with 1.2 nm cutoff as ensemble averages over MD trajectories at ξ values discretized into 0.01 nm bins. The water–water interaction energy was obtained as $U_{WW}(\xi) = H(\xi) - U_{LW}(\xi) - U_{CW}(\xi) - U_{CL}(\xi)$.

The standard formula

$$G(T) = H_0 + C_p(T - T_0) - TS_0 - TC_p \ln \frac{T}{T_0} \quad (2)$$

expresses—for a given system at constant pressure—the temperature dependence of G as a function of the system heat capacity, C_p . Assuming constant C_p for the considered temperature range, such equation can be rewritten in a parametric form as

$$G(T) = A \ln T + BT + C \quad (3)$$

Differentiating with respect to T gives

$$-S(T) = A \ln T + A + B \quad (4)$$

The results presented throughout this article correspond to $S(\xi)$ numerical estimate for $T = 298$ K, obtained from optimal A and B values to fit eq 3 using five $G(\xi)$ values.

Uncertainties on G and S estimates were determined as

$$\delta X = \sqrt{\frac{\sum_{i=1}^N (X_i - \bar{X})^2}{N(N-1)}} \quad (5)$$

where \bar{X} is a value obtained using all data points, while X_i corresponds to $N = 5$ independent simulation blocks. Uncertainties δX for G , H , and $-TS$ profiles along ξ are reported as vertical error bars (Figures 3 and 4).

We note that the uncertainties for entropic and enthalpic components for $T = 298$ K are larger than for $T = 318$ K, the midpoint of the considered temperature range (see the Supporting Information). Nevertheless, as no qualitative difference in the resulting signatures is observed at the two temperatures, we focused on the thermodynamic data for $T = 298$ K in order to remain in line with the majority of other studies conducted at this temperature. We stress that all conclusions of this study regarding the cavity–ligand binding process are independent of this choice.

3. Results

3.1. Hydrophobic Hydration. First, let us focus on the hydration of the ligand and the binding pocket when they are far from each other. Two clearly distinguishable hydration shells are formed around the ligand, in agreement with the expected picture for small-length scale hydration, as inferred from the water density map for $\xi = 1.1$ nm (Figure 2). In contrast, average water density inside the pocket gradually vanishes, indicating that dewetting takes place (Figure 3A). This can be attributed exclusively to the concave topography of pocket walls, as the water potential energy due to interaction with the solutes is lower at the pocket bottom

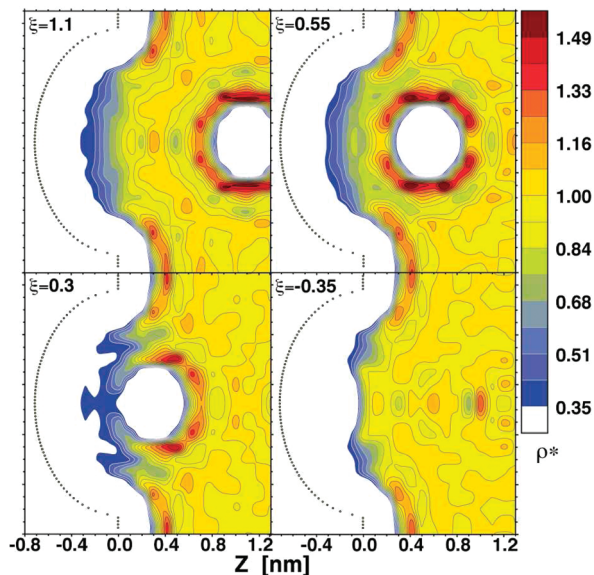


Figure 2. Water density distribution maps for key snapshots along ξ . Color coding is normalized such that $\rho^* = 1$ corresponds to bulk water density of 998 g/L.

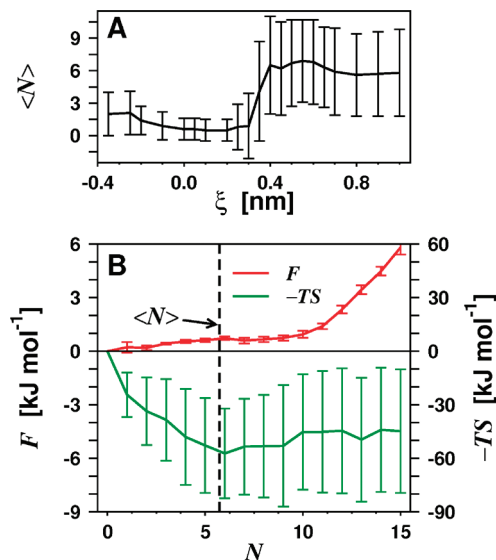


Figure 3. (Thermo)dynamics of pocket hydration. (A) Average cavity occupancy along ξ . $\langle N \rangle$ is the average number of water oxygens located at $Z < 0$. (B) Free energy and its entropy component as functions of pocket occupancy, N (note different energy scale for each profile).

(-4.6 kJ mol $^{-1}$ minimum) than around the ligand (-0.9 kJ mol $^{-1}$ minimum).

A previous analysis of pocket hydration revealed that the observed average water density results from intermittent expansions and retractions of the liquid phase rather than uncorrelated diffusion of individual water molecules characteristic of a “vapor-like” phase.⁵² In order to quantify the thermodynamic effects due to pocket hydration, we calculated the system free energy, F , as a function of pocket occupancy, N , using the relation $F(N) = -k_B T \ln P(N)$, where N is the number of water molecules with an oxygen atom located at $z < 0$, $P(N)$ corresponds to its probability distribution obtained from MD trajectories with the ligand at $\xi \geq 10$ nm, T is the temperature, and k_B is the Boltzmann

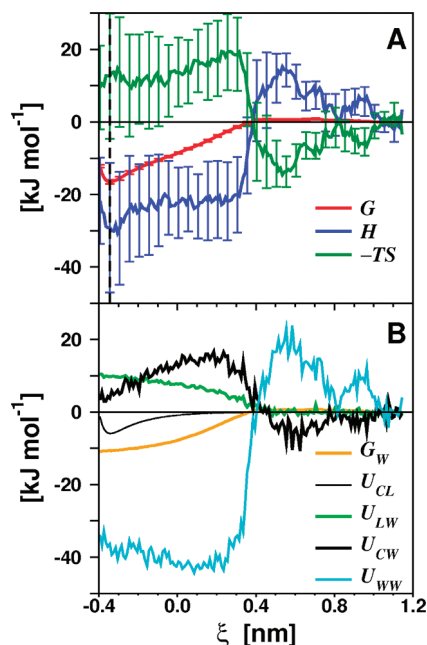


Figure 4. Thermodynamic contributions along the binding coordinate. (A) Relative Gibbs free energy, G (red), enthalpy, H (blue), and entropic term, $-TS$ (green), and their uncertainties (vertical bars; eq 5). (B) Water contribution to the Gibbs free energy, G_w (orange), and decomposed interaction energies: cavity–ligand, U_{CL} (thin black), ligand–water, U_{LW} (green), cavity–water, U_{CW} (black), and water–water, U_{WW} (cyan).

constant. The F profile reveals series of metastable states extending from $N = 0$ (empty pocket) through $N \approx 6$ (average occupancy) to $N = 11$ (close to bulk-like density), indicating no apparent free energy change for pocket dehydration (Figure 3B). The corresponding entropy changes were derived using the temperature dependence of the free energy based on MD simulations at five different temperatures (see Methods Section). They show that complete pocket dehydration, starting from the average water occupancy, is entropically unfavorable. This observation agrees with the fact that solvent fluctuations are eliminated upon dehydration. At the same time, the free energy plateau reveals a perfect entropy–enthalpy compensation, indicating the favorable enthalpic effect due to the displacement of water molecules from the hydrophobic cavity interior to the bulk.

3.2. Cavity–Ligand Association. The obtained free energy profile for the ligand approaching the cavity is flat until $\xi \sim 0.4$ nm. At shorter distances, it decreases monotonically indicating a steady mean force toward binding (Figure 4A). The free energy minimum is reached at $\xi = -0.35$ nm and corresponds to a direct contact between the ligand and the pocket bottom. The overall free-energy change upon binding, $\Delta G = -16.5 \pm 0.6$ kJ mol^{-1} , is dominated by water-related contributions, $\Delta G_w = -10.6 \pm 0.6$ kJ mol^{-1} , as direct cavity–ligand interaction energy ΔU_{CL} is only -5.9 kJ mol^{-1} (Figure 4B). Differently from what is found for other processes driven by hydrophobic interactions, nonpolar cavity–ligand binding is driven here by a large change in enthalpy, $\Delta H = -29.1 \pm 17.3$ kJ mol^{-1} , with an opposing entropic contribution, $-T\Delta S = 12.6 \pm 17.3$ kJ mol^{-1} .

A relatively simple shape of the free energy profile covers substantial, compensating changes in enthalpy and entropy components as the ligand moves toward the cavity. A local maximum in H of 15.2 ± 6.1 kJ mol^{-1} , mirrored by a minimum in $-TS$ of -14.6 ± 6.1 kJ mol^{-1} , is observed already around $\xi = 0.55$ nm. A comparison between water density maps for $\xi = 0.55$ and 1.1 nm (Figure 2) indicates that the reason for the observed changes is the partial destruction of the ligand's second hydration shell, leaving the first hydration shell exposed to the weakly hydrated pocket region. Accordingly, a major contribution to the increasing enthalpy arises from the change in water–water interaction energy (U_{WW}), with marginal effects due to cavity–water (U_{CW}) and ligand–water (U_{LW}) interactions.

The trends of enthalpic and entropic components suddenly invert as the ligand moves closer to the pocket, leading to a substantial decrease in H until $\xi = 0.3$ nm, with a subsequent plateau, and the formation of a $-TS$ maximum (Figure 4A). These changes correspond to almost complete dehydration of the region between the cavity and the ligand (Figures 2 and 3A). The sizable enthalpy change observed is due to favorable contribution from water–water interactions that dominate unfavorable energetic effects of cavity and ligand dehydration. It is worth noting that no simultaneous change in free energy occurs, in agreement with the free energy profile for cavity dehydration described in the previous section. In addition, the change in the entropic component of PMF for the ligand moving from $\xi = 0.55$ to 0.3 nm (34 ± 11 kJ mol^{-1}) is comparable with the entropic effect of changing the cavity occupancy from the average value to $N = 1$ (33 ± 28 kJ mol^{-1}). This agreement might be partially fortuitous and should be treated semiquantitatively because of a number of factors. First, the cavity dehydration upon ligand binding and dewetting due to spontaneous fluctuations are different processes. Second, at $\xi = 0.3$ nm, the dehydration is not fully complete. Third, large uncertainties are associated with $-TS$ values. Yet, the observed correspondence seems to confirm the overall interpretation of the described thermodynamic effects.

Further ligand translocation into the pocket corresponds to the gradual withdrawal of a hydrophobic object from an aqueous environment. It is accompanied by a steady decrease in free energy, initially driven by favorable changes in entropy. At the same time, no apparent change in enthalpy takes place, due to compensation between the increase in U_{LW} and the decrease in U_{CW} . For $\xi < -2.5$ nm, any extra ligand dehydration is limited. Increasingly favorable cavity–water interactions together with increasingly strong direct cavity–ligand attraction lead to a favorable enthalpy change. The moderate increase in the entropy component around the equilibrium binding distance (~ 3 kJ mol^{-1}) can be explained considering ligand immobilization due to direct contact with the wall.

4. Discussion

Our results demonstrate the determinant role of pocket (de)hydration upon hydrophobic cavity–ligand binding on the underlying thermodynamic signature. The favorable, driving enthalpy change results from the release of water

molecules from the hydrophobic environment to bulk water. An opposing entropic component arises due to the elimination of solvent fluctuations inside the pocket. Such contributions dominate over the effects of extracting the small nonpolar ligand out of solution. Thus, cavity–ligand binding is enthalpy, rather than entropy, driven in our system and differs in this respect from other cases of nanoscopic hydrophobic association.^{23,28,33} The hydrophobic association investigated herein is peculiar in that partial dewetting of one of the interacting partners occurs natively, already prior to binding due to the concave cavity geometry.

Remarkably, pocket dehydration displays virtually perfect enthalpy–entropy compensation, i.e., no net contribution to free energy. This is not expected to be a general phenomenon, as the free energy penalty upon dewetting depends on the specific enclosure size and the degree of hydrophobicity.^{26,36,54–57} Larger or more hydrophilic pockets should be preferentially hydrated, while smaller hydrophobic pockets should remain natively dry.

Is our model system unique? A comprehensive study by Young and co-workers³⁴ focused on 14 protein structures selected from the Protein Data Bank (PDB) database for their large hydrophobic binding cavities. In explicit solvent molecular dynamics (MD) simulations drying was found in six cases, sometimes with few (usually two to five) water molecules fluctuating inside the cavity. Most of the reported cavities were long and narrow, as opposed to the hemispherical widely open pocket investigated herein. The driving forces for their hydration remain unknown, leaving open the interesting problem whether the architecture of the cavity may determine qualitative differences in thermodynamic signature for its interaction with water. A comparison of our results with the entropically hampered hydration found in spherical, *closed* cavities³⁶ indicates that the actual topography of the confining potential well and the degree of its openness toward bulk may be important in this respect. On the other hand, in line with our study, entropy-driven hydration was reported for a closed, nonpolar cavity of the I76A mutant of barnase.⁴¹

The (thermo)dynamics of cavity–ligand association addressed herein may depend on even a broader variety of subtle, balancing effects. Recent experimental studies shed light on nonpolar ligand binding to the poorly solvated pocket of the mouse major urinary protein-1 (MUP-1).^{42–44} Despite the apparent hydrophobic character of the binding partners, association appears to be enthalpy-driven and accompanied by an unfavorable entropy change. Following a careful analysis, the authors attribute most of the enthalpic contribution to the effect of water displacement,⁴⁴ and the entropic penalty exclusively to the restriction of ligand and receptor degrees of freedom.⁴³ At the same time, in line with the argument by Dunitz, they assume that the release of bound water molecules is entropically favorable. Interestingly, however, the considered binding site seems to be hydrated by rather disorganized water.⁴⁴ This finding rises an intriguing question: is it legitimate to attribute the observed thermodynamic signature to loss in water entropy? As indicated by our study, such a scenario – in which cavity water is more entropic than bulk water – is possible.

The investigation of model cavity–ligand recognition to systems with varying physicochemical properties is being undertaken by means of the approach presented herein.⁵⁸

Acknowledgment. This work was supported, in part, by the National Institutes of Health, the National Science Foundation, and the Howard Hughes Medical Institute. We thank the Center for Theoretical Biological Physics (NSF Grant PHY-0822283) for the computing resources employed and Dr. Joachim Dzubiella for a critical reading of the manuscript.

Supporting Information Available: Model description for the hydrophobic wall and its interaction with water, and additional thermodynamic data for $T = 318$ K. This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- (1) Lum, K.; Luzzi, A. *Phys. Rev. E: Stat. Phys., Plasmas, Fluids, Relat. Interdiscip. Top.* **1997**, *56*, R6283–R6286.
- (2) Rajamani, S.; Truskett, T. M.; Garde, S. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 9475–9480.
- (3) Ashbaugh, H. S.; Pratt, L. R. *Rev. Mod. Phys.* **2006**, *78*, 156–178.
- (4) Meyer, E. E.; Rosenberg, K. J.; Israelachvili, J. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 15739–15746.
- (5) Chandler, D. *Phys. Rev. E: Stat. Phys., Plasmas, Fluids, Relat. Interdiscip. Top.* **1993**, *48*, 2898–2905.
- (6) Hummer, G.; Garde, S.; Garcia, A. E.; Pohorille, A.; Pratt, L. R. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 8951–8955.
- (7) Hummer, G.; Garde, S.; Garcia, A.; Paulaitis, M.; Pratt, L. J. *Phys. Chem. B* **1998**, *102*, 10469–10482.
- (8) Chandler, D. *Nature* **2005**, *437*, 640–647.
- (9) Smith, D. E.; Zhang, L.; Haymet, A. D. J. *J. Am. Chem. Soc.* **1992**, *114*, 5875–5876.
- (10) Smith, D. E.; Haymet, A. D. J. *J. Chem. Phys.* **1993**, *98*, 6445–6454.
- (11) Ludemann, S.; Abseher, R.; Schreiber, H.; Steinhauser, O. *J. Am. Chem. Soc.* **1997**, *119*, 4206–4213.
- (12) Shimizu, S.; Chan, H. S. *J. Chem. Phys.* **2000**, *113*, 4683–4700.
- (13) Raschke, T. M.; Tsai, J.; Levitt, M. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 5965–5969.
- (14) Lee, C. Y.; McCammon, J. A.; Rossky, P. J. *J. Chem. Phys.* **1984**, *80*, 4448–4455.
- (15) Berne, B. J.; Weeks, J. D.; Zhou, R. *Annu. Rev. Phys. Chem.* **2009**, *60*, 85–103.
- (16) Hummer, G.; Rasaiah, J. C.; Noworyta, J. P. *Nature* **2001**, *414*, 188–190.
- (17) Choudhury, N.; Pettitt, B. *J. Am. Chem. Soc.* **2007**, *129*, 4847–4852.
- (18) Stillinger, F. H. *J. Solution Chem.* **1973**, *2*, 141–158.
- (19) Wallqvist, A.; Gallicchio, E.; Levy, R. M. *J. Phys. Chem. B* **2001**, *105*, 6745–6753.
- (20) Jensen, T. R.; Jensen, M. O.; Reitzel, N.; Balashev, K.; Peters, G. H.; Kjaer, K.; Bjornholm, T. *Phys. Rev. Lett.* **2003**, *90*, 086101–1–086101–4.

- (21) Young, T.; Abel, R.; Kim, B.; Berne, B. J.; Friesner, R. A. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 808–813.
- (22) Qvist, J.; Davidovic, M.; Hamelberg, D.; Halle, B. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 6296–6301.
- (23) Wallqvist, A.; Berne, B. J. *J. Phys. Chem.* **1995**, *99*, 2893–2899.
- (24) ten Wolde, P. R.; Chandler, D. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 6539–6543.
- (25) Huang, X.; Margulis, C. J.; Berne, B. J. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 11953–11958.
- (26) Liu, P.; Huang, X.; Zhou, R.; Berne, B. J. *Nature* **2005**, *437*, 159–162.
- (27) Choudhury, N.; Pettitt, B. M. *J. Am. Chem. Soc.* **2005**, *127*, 3556–3567.
- (28) Choudhury, N.; Pettitt, B. M. *J. Phys. Chem. B* **2006**, *110*, 8459–8463.
- (29) Athawale, M. V.; Goel, G.; Ghosh, T.; Truskett, T. M.; Garde, S. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 733–738.
- (30) Setny, P. *J. Chem. Phys.* **2007**, *127*, 054505.
- (31) Hua, L.; Huang, X.; Liu, P.; Zhou, R.; Berne, B. J. *J. Phys. Chem. B* **2007**, *111*, 9069–9077.
- (32) Willard, A. P.; Chandler, D. *J. Phys. Chem. B* **2008**, *112*, 6187–6192.
- (33) Zangi, R.; Berne, B. J. *J. Phys. Chem. B* **2008**, *112*, 8634–8644.
- (34) Young, T.; Hua, L.; Huang, X.; Abel, R.; Friesner, R.; Berne, B. J. *Proteins: Struct., Funct., Bioinf.* **2010**, *78*, 1856–1869.
- (35) Dunitz, J. D. *Science* **1994**, *264*, 670.
- (36) Vaitheeswaran, S.; Yin, H.; Rasaiah, J. C.; Hummer, G. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 17002–17005.
- (37) Yin, H.; Hummer, G.; Rasaiah, J. C. *J. Am. Chem. Soc.* **2007**, *129*, 7369–7377.
- (38) Rasaiah, J. C.; Garde, S.; Hummer, G. *Annu. Rev. Phys. Chem.* **2008**, *59*, 713–740.
- (39) Ernst, J. A.; Clubb, R. T.; Zhou, H. X.; Gronenborn, A. M.; Clore, G. M. *Science* **1995**, *267*, 1813–1817.
- (40) Denisov, V. P.; Venu, K.; Peters, J.; Horlein, H. D.; Halle, B. *J. Phys. Chem. B* **1997**, *101*, 9380–9389.
- (41) Olano, L. R.; Rick, S. W. *J. Am. Chem. Soc.* **2004**, *126*, 7991–8000.
- (42) Sharrow, S. D.; Novotny, M. V.; Stone, M. J. *Biochemistry (Moscow)* **2003**, *42*, 6302–6309.
- (43) Bingham, R. J.; Findlay, J. B. C.; Hsieh, S.-Y.; Kalverda, A. P.; Kjellberg, A.; Perazzolo, C.; Phillips, S. E. V.; Seshadri, K.; Trinh, C. H.; Turnbull, W. B.; Bodenhausen, G.; Homans, S. W. *J. Am. Chem. Soc.* **2004**, *126*, 1675–1681.
- (44) Barratt, E.; Bingham, R. J.; Warner, D. J.; Laughton, C. A.; Phillips, S. E. V.; Homans, S. W. *J. Am. Chem. Soc.* **2005**, *127*, 11827–11834.
- (45) Setny, P.; Wang, Z.; Cheng, L.-T.; Li, B.; McCammon, J. A.; Dzubiella, J. *Phys. Rev. Lett.* **2009**, *103*, 187801.
- (46) Jorgensen, W. L.; Madura, J. D.; Swenson, C. J. *J. Am. Chem. Soc.* **1984**, *106*, 6638–6646.
- (47) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926–935.
- (48) Torrie, G.; Valleau, J. J. *Comput. Phys.* **1977**, *23*, 187–199.
- (49) Kumar, S.; Rosenberg, J. M.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A. *J. Comput. Chem.* **1992**, *13*, 1011–1021.
- (50) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* **1983**, *4*, 187–217.
- (51) Cooper, J. *Revised Release on the IAPWS Industrial Formulation 1997 for the Thermodynamic Properties of Water and Steam*; University of London: London; 2007.
- (52) Setny, P.; Geller, M. *J. Chem. Phys.* **2006**, *125*, 144717.
- (53) Preusser, A. *ACM Trans. Math. Software* **1998**, *15*, 79–89.
- (54) Brovchenko, I.; Paschek, D.; Geiger, A. *J. Chem. Phys.* **2000**, *113*, 5026–5036.
- (55) Li, Z.; Lazaridis, T. *Phys. Chem. Chem. Phys.* **2007**, *9*, 573–81.
- (56) Giovambattista, N.; Lopez, C. F.; Rosky, P. J.; Debenedetti, P. G. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 2274–2279.
- (57) Carey, C.; Cheng, Y.-K.; Rosky, P. *J. Chem. Phys.* **2000**, *258*, 415–425.
- (58) Baron, R.; Setny, P.; McCammon, J. A. *J. Am. Chem. Soc.* **2010**, *132*, 12091–12097.

CT1003077

Computational Thermochemistry: Scale Factor Databases and Scale Factors for Vibrational Frequencies Obtained from Electronic Model Chemistries

I. M. Alecu,[†] Jingjing Zheng,[†] Yan Zhao,[‡] and Donald G. Truhlar^{*†}

Department of Chemistry and Supercomputing Institute, University of Minnesota, Minneapolis, Minnesota 55455-0431 and Commercial Print Engine Lab, HP Laboratories, Hewlett-Packard Company, Palo Alto, California 94304

Received June 15, 2010

Abstract: Optimized scale factors for calculating vibrational harmonic and fundamental frequencies and zero-point energies have been determined for 145 electronic model chemistries, including 119 based on approximate functionals depending on occupied orbitals, 19 based on single-level wave function theory, three based on the neglect-of-diatom-differential-overlap, two based on doubly hybrid density functional theory, and two based on multicoefficient correlation methods. Forty of the scale factors are obtained from large databases, which are also used to derive two universal scale factor ratios that can be used to interconvert between scale factors optimized for various properties, enabling the derivation of three key scale factors at the effort of optimizing only one of them. A reduced scale factor optimization model is formulated in order to further reduce the cost of optimizing scale factors, and the reduced model is illustrated by using it to obtain 105 additional scale factors. Using root-mean-square errors from the values in the large databases, we find that scaling reduces errors in zero-point energies by a factor of 2.3 and errors in fundamental vibrational frequencies by a factor of 3.0, but it reduces errors in harmonic vibrational frequencies by only a factor of 1.3. It is shown that, upon scaling, the balanced multicoefficient correlation method based on coupled cluster theory with single and double excitations (BMC–CCSD) can lead to very accurate predictions of vibrational frequencies. With a polarized, minimally augmented basis set, the density functionals with zero-point energy scale factors closest to unity are MPWLYP1M (1.009), τ HCTHhyb (0.989), BB95 (1.012), BLYP (1.013), BP86 (1.014), B3LYP (0.986), MPW3LYP (0.986), and VSXC (0.986).

1. Introduction

The accurate determination of vibrational frequencies through the use of computational quantum chemistry is essential to many fields of chemistry. For example, computed frequencies can be used to guide spectroscopic measurements by predicting or refining the spectral regions in which transitions of interest might occur, and computed frequencies can also help to identify and characterize transient species, such as molecular radicals, van der Waals complexes, and reactive

intermediates. Furthermore, frequencies are an essential part of computational thermochemistry, and they are used for the calculation of vibrational zero-point energies (ZPEs) and vibrational partition functions, which are also important for thermochemical kinetics.

One goal of computational quantum chemistry is to attain chemical accuracy, which is generally defined for energetic quantities as an accuracy of 1 kcal mol⁻¹. The development of increasingly precise electronic model chemistries coupled with improved computational efficiency can now consistently permit the calculation of even more accurate energies (e.g., 1 kJ mol⁻¹) for small systems. In fact, Born–Oppenheimer energies can be calculated to such a degree of accuracy by

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

[†] University of Minnesota.

[‡] Hewlett-Packard Company.

some modern electronic model chemistries that the principal source of disagreement between these computations and experiment can shift from unconverged electronic energies to the approximate treatment of ZPEs, thermal vibrational energies, and vibrational free energies. Consequently, the calculation of accurate vibrational frequencies is crucial. For example, the ZPE of 1-pentanol is slightly in excess of 100 kcal mol⁻¹, so a 1% error in the ZPE is too large if one is aiming for chemical accuracy.

Electronic model chemistries are usually used to compute vibrational frequencies by invoking the approximation that the potential energy in the vicinity of a minimum varies quadratically with respect to the nuclear coordinates. For ZPEs, this leads to the expression

$$\varepsilon_{\text{vib}}^{\text{G}} \cong \varepsilon_{\text{vib}}^{\text{GH}} = \frac{\hbar c}{2} \sum_m \omega_m \quad (1)$$

where $\varepsilon_{\text{vib}}^{\text{G}}$ is the ZPE (difference between ground vibrational-state energy and the potential energy at the classical equilibrium position), $\varepsilon_{\text{vib}}^{\text{GH}}$ is the harmonic approximation to it, ω_m is the computed harmonic vibrational frequency of mode m in wave numbers (this is computed from the Hessian, that is, the matrix of quadratic force constants), \hbar is Planck's constant, and c is the speed of light. The neglect of anharmonicity tends to overestimate the ZPE.¹ Therefore, in order to obtain accurate zero-point energies, one must either calculate the anharmonic corrections or scale the frequencies. Calculating the anharmonicity requires higher-order (e.g., cubic and quartic) force constants and information about torsional barriers or full potential energy surfaces, and this additional data is often unavailable or even unaffordable. For small molecules, perturbative approaches may be affordable and can yield accurate anharmonic vibrational frequencies provided that these approaches are not significantly affected by the problem of Fermi resonances.^{2–6} However, linear species cannot be treated with this formalism, and for large systems, the associated computational cost may not be affordable since analytic second derivatives are required for such calculations. Therefore, in order to obtain more accurate ZPEs directly from the computed quadratic force constants for complex systems, it is customary to use scaling, and this is the subject of the present article.

We will scale the harmonic frequencies by a constant:

$$\varepsilon_{\text{vib}}^{\text{G}} \cong \frac{\hbar c}{2} \sum_m (\lambda \omega_m) \quad (2)$$

Alternatively, it might be of interest to obtain the true harmonic or fundamental frequencies from the set of computed harmonic frequencies:

$$\omega_m^{\text{true}} \cong \lambda \omega_m \quad (3)$$

$$v_m \cong \lambda \omega_m \quad (4)$$

where ω_m^{true} are the true harmonic frequencies, and v_m are the observable fundamental frequencies. The constant λ is commonly referred to as a “scale factor,” and its specific value depends on the molecule and the electronic model chemistry used to compute the frequencies. We will use the

language of Pople,⁷ by which an electronic structure method is called a theoretical model chemistry or—in the present context—an electronic model chemistry, and for single-level methods, it denotes a choice of one-electron basis set and either a level of correlation of the electronic wave function or a choice of approximate density functional. An electronic model chemistry can also be a multilevel method, such as BMC-CCSD⁸ or MC3BB.⁹

For highest accuracy, the values of the scale factors λ will be different in each of eqs 2–4, i.e., they will depend not only on the electronic model chemistry but also on the property for which they are optimized.^{10,11} In the present work, we optimize general scale factors that can be used to obtain accurate fundamental (λ^{F}) and true harmonic frequencies (λ^{H}) and ZPEs (λ^{ZPE}). These optimized values are obtained, for a given electronic model chemistry, by correlating computed harmonic frequencies with specially constructed experimental databases. In addition, we also test the assumption that the scale factors λ^{F} , λ^{H} , and λ^{ZPE} can be related to one another through simple proportionality constants. Accepting these proportionalities allows us to derive a set of universal scale factor ratios that can be used to interconvert between scale factors optimized for different purposes, making it possible to obtain all three scale factors at the expense of explicitly calculating just one, for any given electronic model chemistry. Finally, we construct a representative database of accurate zero-point energies that is smaller than the original one but statistically representative, so that additional scale factor optimizations can be carried out with minimal effort, allowing—in the future—the convenient derivation of scale factors for new density functionals, new wave function approximations, and new basis sets.

2. Methodology

2.1. Accurate Vibrational Frequencies and Zero-Point Energies. In order to determine the optimal values for the scale factors of interest, it is essential to first establish reliable databases of accurate experimental quantities to which the computed analogs can be compared. We recall that although fundamental frequencies can be observed directly, the corresponding harmonic analogs are just conceptual, and in practice, these are obtained from experiment by extrapolating a vibrational progression to the experimentally inaccessible vibrationless state.¹² For harmonic frequencies, we employ the recently compiled F38/06 database,¹³ which consists of the 38 experimental harmonic frequencies of the 15 molecules used in this study, originally taken from Martin et al.¹⁴ (Note that F in the name of the database denotes frequencies, although elsewhere in the article it denotes fundamentals.) In Table 1 we supplement the F38/06 database with an additional column containing the 38 experimentally observed fundamental frequencies for the 15 molecules in question, taken from the Computational Chemistry Comparison and Benchmark Database and the NIST Chemistry Web book.^{15,16} This compilation of the harmonic and fundamental frequencies will be denoted as the F38/10 database.

Table 1. F38/10 Database: Experimental Harmonic and Fundamental Vibrational Frequencies

molecule	mode	symmetry	harmonic frequency ^a (cm ⁻¹)	fundamental frequency ^b (cm ⁻¹)
HF	1	σ	4138	3959
H ₂	1	σ_g	4401	4159
N ₂	1	σ_g	2359	2330
F ₂	1	σ_g	917	894
CO	1	σ	2170	2143
OH	1	σ	3738 ^b	3568
Cl ₂	1	σ_g	560 ^b	554
CO ₂	1	σ_g	1354	1333
	2	σ_u	2397	2349
	3	π_u	673	667
H ₂ O	1	a_1	3832	3657
	2	a_1	1648	1595
	3	b_1	3942	3756
N ₂ O	1	σ	2282	2224
	2	σ	1298	1285
	3	π	596	589
HCN	1	σ	3443	3312
	2	σ	2127	2089
	3	π	727	712
C ₂ H ₂	1	σ_g	3495	3374
	2	σ_g	2008	1974
	3	σ_u	3415	3289
	4	π_g	624	612
	5	π_u	747	730
H ₂ CO	1	a_1	2937	2782
	2	a_1	1778	1746
	3	a_1	1544	1500
	4	b_1	1188	1167
	5	b_2	3012	2843
	6	b_2	1269	1249
NH ₃	1	a_1	3478	3337
	2	a_1	1084	950
	3	e	3597	3444
	4	e	1684	1627
CH ₄	1	a_1	3026	2917
	2	e	1583	1534
	3	t_2	3157	3019
	4	t_2	1367	1306

^a Martin et al.¹⁴ ^b NIST Web book and Computational Chemistry Comparison and Benchmark Database.^{15,16}

Because real molecules cannot have stationary energies lower than those of their ground states, the ZPE, which in quantum chemistry is defined as the difference between the energy of the vibrational ground state and the energy minimum on the Born–Oppenheimer¹⁷ or Born–Huang¹⁸ potential energy surface, is not a measurable quantity. However, ZPEs have been estimated from experiment by a two-step process in which first the spectroscopic data are used to infer a potential energy surface or a set of spectroscopic constants, and then the ZPE is inferred from the surface or constants. In previous work¹³ on the optimization of scale factors for ZPEs, we have used the ZPVE15/06 database, which is composed of such “experimental” ZPEs, originally taken from the compilations of Martin¹⁹ and Grev et al.,¹⁰ for the 15 molecules investigated here. Recently, however, Irikura and co-workers have meticulously reanalyzed the available spectroscopic parameters for many important molecules, including 12 of the 15 used in this study and arrived at somewhat different values for their respective ZPEs.^{20,21} In addition, Irikura and co-workers have also carefully taken into account the propagation of the statistical

uncertainties associated with the spectroscopic constants reported in literature as well as those resulting from bias, to arrive at an estimate of the standard uncertainty for the ZPEs derived in their work.^{20,21}

Aside from the aim of compiling a list of precise ZPEs against which computational methods can be validated, Irikura’s reanalysis of these quantities was further motivated by the fact that the formulas commonly used to extract “experimental” ZPEs from spectroscopic constants were often inadequate due to the inclusion of only a subset of terms in the power series of vibrational energy as a function of vibrational quantum numbers.^{2,4,20–26} For example, in the case of a diatomic molecule, the most commonly used formula for obtaining the vibrational energy levels relative to the energetic minimum on the potential energy curve is

$$G(v) \cong \omega_e \left(v + \frac{1}{2} \right) - \omega_e x_e \left(v + \frac{1}{2} \right)^2 + \omega_e y_e \left(v + \frac{1}{2} \right)^3 + \dots \quad (5)$$

where ω_e , $\omega_e x_e$, and $\omega_e y_e$ denote parameters corresponding to the harmonic frequency and the first and second anharmonicity constants, respectively, and v is the vibrational quantum number.²⁷ This leads to the expression:

$$\frac{\epsilon_{\text{vib}}^G}{hc} \equiv G(0) \cong \frac{1}{2}\omega_e - \frac{1}{4}\omega_e x_e + \frac{1}{8}\omega_e y_e \quad (6)$$

for the ZPE. In contrast, Dunham has shown²⁸ that the ZPE of a diatomic species can be expressed more accurately through the power series:

$$G(v) \cong Y_{00} + Y_{10} \left(v + \frac{1}{2} \right) - Y_{20} \left(v + \frac{1}{2} \right)^2 + Y_{30} \left(v + \frac{1}{2} \right)^3 + \dots \quad (7)$$

where Y_{10} , Y_{20} , and Y_{30} approximately correspond to ω_e , $\omega_e x_e$, and $\omega_e y_e$, and Y_{00} is a constant that is commonly neglected. The values of the constant term Y_{00} and the higher-order anharmonicity $\omega_e y_e$ are often unknown and even more often neglected; however, although they are usually small (often no more than a few cm⁻¹), they are not completely negligible and taking them into account leads to more accurate estimations of the diatomic ZPEs.²⁰ The inclusion of all of these constants into eq 6 leads to

$$\frac{\epsilon_{\text{vib}}^G}{hc} \equiv G(0) \cong Y_{00} + \frac{1}{2}\omega_e - \frac{1}{4}\omega_e x_e + \frac{1}{8}\omega_e y_e \quad (8)$$

which is the formula used by Irikura to derive the ZPEs for the diatomics in his work.²⁰

A constant term called E_0 ,^{2,22} analogous to Y_{00} , has often been neglected in analyzing polyatomic spectra. Irikura et al. have estimated this term and higher-order anharmonic terms for polyatomics and thereby also derived refined values for polyatomic ZPEs.²¹ Therefore, in the present work, we update the original ZPVE15/06 database by replacing the old ZPE values from this database with the recently refined values from Irikura and co-workers, with the exception of N₂O, NH₃, and CH₄, whose ZPE values have not been reanalyzed by Irikura et al.^{20,21} We denote this new compila-

Table 2. ZPVE15/10 Database: Experimental ZPEs for the 15 Molecules Used in This Work

molecule	ZPE ^a (kcal mol ⁻¹)
HF	5.86353 ^a
H ₂	6.2310 ^a
N ₂	3.3618 ^a
F ₂	1.3021 ^a
CO	3.0929144 ^a
OH	5.2915 ^a
Cl ₂	0.7983 ^a
CO ₂	7.3 ^a
H ₂ O	13.26 ^a
N ₂ O	6.770 ^b
HCN	10.0 ^a
C ₂ H ₂	16.49 ^a
H ₂ CO	16.1 ^a
NH ₃	21.200 ^c
CH ₄	27.710 ^b

^a Irikura et al.^{20,21} ^b Grev et al.¹⁰ ^c Martin.¹⁹

tion of 15 ZPEs as the ZPVE15/10 database, and we present it in Table 2.

2.2. Computational Methods. In the first stage of this work, the harmonic frequencies of the 15 molecules comprising the F38/10 database have been computed using 22 combinations of approximate density functionals with one-electron basis sets and also using the BMC-CCSD⁸ multicoefficient correlation method. The density functionals used in this first stage are B1LYP,^{29–31} M05-2X,³² M06-L,³³ M06-HF,³⁴ M06,¹³ M06-2X,¹³ M08-SO,³⁵ and M08-HX.³⁵ Each of these functionals was paired with both the 6-31+G(d,p)^{1,36} double- ζ basis sets and several triple- ζ basis sets: MG3S,³⁷ def2-TZVPP,^{38,39} MG3SXP,³⁵ cc-pVTZ+,^{40–43} aug-cc-pVTZ,^{40,42,44} and maug-cc-pV(T+d)Z.^{40,42–45} In this stage, we also reoptimized the ZPE scale factors obtained with HF/MG3S theory, where HF denotes Hartree-Fock, and 16 important functional/MG3S combinations that have recently been published,¹³ and we supplemented these with newly derived scale factors for frequencies. The 16 functionals involved in this step are BLYP,^{29,30} B3LYP,^{29,30,46,47} PBE,⁴⁸ B98,⁴⁹ VSXC,⁵⁰ PBE0,^{51–53} HFLYP,^{54,55} TPSSH,⁵⁶ BMK,⁵⁷ B97-3,⁵⁸ M05,⁵⁹ M05-2X,³² M06-L,³³ M06-HF,³⁴ M06,¹³ and M06-2X.¹³ In all, the set of three scale factors were optimized for 40 electronic model chemistries using the full databases in stage one.

In the second stage of this study, we obtain and employ a representative database (which is a key component of what will be called a reduced optimization model) or, in some cases, make use of approximate systematic corrections to efficiently derive and/or update scale factors for 105 additional electronic model chemistries. The electronic model chemistries used in stage two are some of the density functionals listed above, plus single-level wave function approximations: HF, MP2,¹ MP4(SDQ),¹ QCISD,⁶⁰ CCSD,⁶¹ CCSD(T),⁶² CCSD-F12,^{63,64} CCSD(T)-F12,^{63,64} other density functionals that depend on occupied orbitals: B1B95,^{30,65} B3P86,^{46,66} B3PW91,^{46,67–71} BB1K,^{30,65,72} BB95,^{30,65} BP86,^{30,66} CAM-B3LYP,⁷³ HSEh1PBE,^{74–80} MOHLYP,^{29,81} MPW1B95,^{65,82,83} MPW1K,^{83–85} MPW3LYP,^{29,82,83,86} MPWB1K,^{65,82,83} MPWL1P1M,^{29,81} PBE1K CIS,^{48,87,88} PW6B95,⁸⁹ SOGGA,⁹⁰ τ -HCTHhyb,⁸⁴ TPSS1K CIS,^{87,91,92}

ω B97,⁹³ ω B97X,⁹³ XB1K,^{29,30,65,67,82,86} X1B95,^{29,30,65,67,82,86} one of these reoptimized with molecular mechanics dispersion terms: ω B97X-D,⁹⁴ three methods based on the neglect of diatomic differential overlap: AM1,⁹⁵ PM3,^{96,97} PM6,⁹⁸ two doubly hybrid density functional methods: MC3BB,⁹ MC3MPW,⁹ and one more multicoefficient correlation method: MC-QCISD/3.⁹⁹ The density functionals and single-level wave function approximations are paired with various basis sets, many of which have also been used in stage one, plus def2-TZVP,^{38,39} ma-TZVP,^{38,39,100} jul-cc-pVTZ,^{40,42–45,101} MIDI! (also called MIDIX),^{102,103} MIDIY,^{102–104} and a host of Pople-type basis sets.

The selection of these electronic model chemistries was made to strategically complement and extend our pre-existing database¹⁰⁵ of available scale factors for electronic model chemistries, a copy of which is provided in the form of Table S1 in the Supporting Information. All of the density functional calculations employed “ultrafine” integration grids, and, unless otherwise indicated, were carried out with a locally modified version of the Gaussian 03 program suite.^{106,107} The BMC-CCSD, MC3BB, MC3MPW, and MC-QCISD/3 calculations were performed using the ML-GAUSS program.¹⁰⁸

A comment on basis sets may be helpful here. We have shown^{100,101} that a very suitable basis set for many density functional applications is a polarized valence-triple- ζ basis set minimally augmented with diffuse functions, where minimal augmentation is defined as adding diffuse s and p subshells to atoms heavier than hydrogen or helium. Several of the basis sets used here correspond to this “minimally augmented polarized triple- ζ ” class, in particular, MG3S,³⁷ MG3SXP,³⁵ cc-pVTZ+,^{40–43} maug-cc-pV(T+d)Z,^{40,42–45} and ma-TZVP;^{38,39,100} full details of these basis sets and all the other basis sets used here are given in the references cited.

2.3. Scale Factors for Vibrational Frequencies and Zero-Point Energies. To determine the optimal scale factors for the true harmonic and fundamental frequencies, the computed harmonic frequencies were fit against the experimental values available for these quantities, found in the F38/10 database (Table 1). The scale factors λ^F and λ^H were optimized by minimizing the respective root-mean-square (rms) deviations for the entire set of 38 frequencies. Specifically, the rms deviations for the two properties are defined as

$$\text{RMS}(\text{harmonic frequencies}) = \left\{ \left[\sum_{m=1}^{38} (\lambda^H \omega_m - \omega_m^{\text{true}})^2 / 38 \right]^{1/2} \right\} \quad (9)$$

$$\text{RMS}(\text{fundamentals}) = \left\{ \left[\sum_{m=1}^{38} (\lambda^F \omega_m - v_m)^2 / 38 \right]^{1/2} \right\} \quad (10)$$

where ω_m is the computed harmonic frequency, ω_m^{true} is the harmonic frequency derived from extrapolations of the observed vibrational progressions, and v_m is the experimental fundamental frequency, all found in Table 1.

Table 3. Compilation of Optimal Scale Factors

model chemistry	λ^{ZPE}	rms deviation ^a		λ^{H}	rms deviation ^b		λ^{F}	rms deviation ^b	
		scaling	no scaling		scaling	no scaling		scaling	no scaling
B1LYP/MG3S	0.978	0.13	0.31	0.994	42	44	0.955	31	117
B3LYP/MG3S	0.983	0.12	0.25	0.998	41	41	0.960	30	104
B97-3/MG3S	0.972	0.14	0.38	0.986	44	57	0.947	35	138
B98/MG3S	0.982	0.12	0.25	0.995	41	42	0.956	34	114
BLYP/MG3S	1.013	0.11	0.19	1.031	38	85	0.991	38	43
BMC-CCSD	0.985	0.14	0.23	1.001	16	16	0.962	34	101
BMK/MG3S	0.971	0.16	0.40	0.984	54	69	0.945	46	147
HF/MG3S	0.919	0.28	1.12	0.932	86	201	0.895	70	288
HFLYP/MG3S	0.899	0.30	1.40	0.912	92	259	0.876	75	347
M05-2X/6-31+G(d,p)	0.961	0.19	0.54	0.974	61	90	0.936	50	171
M05-2X/def2-TZVPP	0.962	0.22	0.52	0.976	63	87	0.938	51	166
M05-2X/MG3S	0.962	0.21	0.53	0.975	60	87	0.937	50	168
M05/MG3S	0.977	0.16	0.33	0.989	56	63	0.951	46	132
M06-2X/6-31+G(d,p)	0.967	0.18	0.45	0.979	60	81	0.940	51	159
M06-2X/aug-cc-pVTZ	0.971	0.20	0.41	0.985	61	72	0.946	48	144
M06-2X/def2-TZVPP	0.970	0.20	0.43	0.983	61	75	0.945	48	148
M06-2X/maug-cc-pV(T+d)Z	0.971	0.20	0.42	0.984	60	73	0.945	48	146
M06-2X/MG3S	0.970	0.19	0.43	0.982	60	74	0.944	47	149
M06-HF/6-31+G(d,p)	0.954	0.29	0.65	0.969	88	119	0.931	77	193
M06-HF/def2-TZVPP	0.958	0.29	0.61	0.970	87	117	0.932	76	191
M06-HF/MG3S	0.955	0.30	0.65	0.967	86	119	0.930	77	194
M06-L/6-31+G(d,p)	0.978	0.10	0.30	0.992	42	46	0.953	35	123
M06-L/def2-TZVPP	0.976	0.11	0.32	0.995	46	48	0.956	33	114
M06-L/MG3S	0.978	0.10	0.30	0.996	45	46	0.958	33	111
M06/6-31+G(d,p)	0.980	0.15	0.29	0.989	56	62	0.950	46	133
M06/def2-TZVPP	0.979	0.19	0.33	0.992	65	68	0.953	48	127
M06/MG3S	0.981	0.18	0.30	0.994	60	68	0.955	49	123
M08-HX/6-31+G(d,p)	0.972	0.18	0.40	0.983	64	78	0.944	51	150
M08-HX/cc-pVTZ+	0.974	0.20	0.39	0.985	66	76	0.946	51	144
M08-HX/def2-TZVPP	0.973	0.20	0.40	0.984	67	78	0.945	52	147
M08-HX/MG3S	0.973	0.19	0.39	0.984	65	76	0.946	50	145
M08-SO/6-31+G(d,p)	0.979	0.18	0.32	0.989	60	66	0.951	49	134
M08-SO/cc-pVTZ+	0.982	0.20	0.30	0.995	64	65	0.956	47	120
M08-SO/def2-TZVPP	0.980	0.21	0.32	0.993	65	68	0.954	48	125
M08-SO/MG3S	0.983	0.20	0.29	0.995	64	66	0.956	48	119
M08-SO/MG3SXP	0.984	0.21	0.29	0.996	66	67	0.957	49	119
PBE/MG3S	1.010	0.09	0.15	1.025	37	72	0.985	33	48
PBE0/MG3S	0.975	0.15	0.34	0.989	50	57	0.950	37	131
TPSSH/MG3S	0.984	0.12	0.23	1.002	37	38	0.963	33	97
V5XC/MG3S	0.986	0.08	0.19	1.001	32	33	0.962	32	98
average ^c		0.18	0.41		58	76		47	142

^a In units of kcal mol⁻¹. ^b In units of cm⁻¹. ^c The average absolute deviation for ZPE is 0.14 kcal mol⁻¹, and the absolute percent deviation is 2.8% (see Supporting Information).

Optimal scale factors for ZPEs (λ^{ZPE}) can be obtained in an analogous manner, by minimizing the rms deviations between the set of the computed harmonic ZPEs for the 15 species and their “experimental” ZPEs:

$$\text{RMS}(\text{ZPE}) = \left\{ \left[\sum_{m=1}^{15} (\lambda^{\text{ZPE}} \varepsilon_{\text{vib}_m}^{\text{GH}} - \varepsilon_{\text{vib}_m}^{\text{G}})^2 / 15 \right]^{1/2} \right\} \quad (11)$$

The best estimates for the “experimental” ZPEs of the 15 species of interest, $\varepsilon_{\text{vib}_m}^{\text{G}}$, are taken from the ZPVE15/10 database (Table 2).

This procedure yields

$$\lambda^{\text{ZPE}} = \frac{\sum_{m=1}^M (\varepsilon_{\text{vib}_m}^{\text{GH}} \varepsilon_{\text{vib}_m}^{\text{G}})}{\sum_{m=1}^M (\varepsilon_{\text{vib}_m}^{\text{GH}})^2} \quad (12)$$

where M is the number of ZPEs. Analogous expressions can be obtained for λ^{H} and λ^{F} by substituting the corresponding experimental and calculated quantities in the above equation and performing the two summations over the number of frequencies with $M = 38$.

3. Results and Discussion

3.1. Stage One: Scale Factors Obtained from Large Databases. 3.1.1. Scale Factors for Vibrational Frequencies.

The optimal scale factors for reproducing the set of vibrational frequencies of interest from the harmonic vibrational frequencies computed with the electronic model chemistries used in stage one of this work are presented in Table 3. As can be seen in Table 3, the rms deviations for the prediction of vibrational frequencies using the optimal scale factors are fairly reasonable, averaging 58 and 47 cm⁻¹ in the case of harmonic and fundamental frequencies, respectively, which mark substantial improvements over the 76 and 142 cm⁻¹ rms deviations that are obtained for the same quantities in the absence of scaling. Overall, upon scaling, the most reliable results are obtained with the BMC-CCSD multilevel method, which has an average rms deviation (average of eqs 9 and 10) of only 20 cm⁻¹. The best performance of any density functional tested is for VSXC, which has an average rms deviation of 32 cm⁻¹, and the best performance of any Minnesota functional is for M06-L, which has an average rms deviation of 39–40 cm⁻¹, depending on the basis set used.

Most of the electronic model chemistries, however, used in conjunction with the appropriate scale factors, yield

frequencies that are usually in reasonable agreement with experiment, with average absolute deviations of $44 \pm 13 \text{ cm}^{-1}$ ($2.9 \pm 0.9\%$) and $35 \pm 10 \text{ cm}^{-1}$ ($2.4 \pm 0.6\%$) for harmonic and fundamental frequencies, respectively; although HF, HFLYP, and M06-HF (all with 100% Hartree–Fock exchange) are much less reliable than the other electronic model chemistries tested (Table 4). However, because the rms deviations can be misleading when trying to select an appropriate method for a specific problem, we have tabulated in Table 4 the number N of “outlier” frequencies for each of the electronic model chemistries tested. We have arbitrarily defined outliers as scaled frequencies that differ from their corresponding experimental values by more than 5%. In addition, Table 4 also contains a list of the outlying frequencies. In an earlier study of computational harmonic frequencies and equilibrium geometries, Martin noted that “the F_2 molecule proves troublesome for all functionals with regard to bond length and harmonic frequencies (this is less the case for HF).”¹⁴ It is interesting to note that even upon scaling, almost all of the electronic model chemistries tested severely overestimated the harmonic and fundamental stretching frequency in F_2 . In fact, only the BMC–CCSD and M06-L/6-31+G(d,p) electronic model chemistries were successful in predicting both the F_2 harmonic and fundamental frequencies with deviations of less than 5%.

Increasing the size of the one-electron basis set for a given density functional did not significantly alter the ensuing values for λ^{F} and λ^{H} nor lead to more accurate results. In fact, increasing the basis set generally resulted in a slight loss of accuracy with regard to the prediction of both the harmonic and fundamental frequencies. We also note that while λ^{F} and λ^{H} change somewhat in going from a double- to a triple- ζ basis set, these parameters do not show much sensitivity toward the nature of the triple- ζ basis set used, with all differences in optimum scale factors being 0.004 or less. These results are consistent with the observations of Andersson and Uvdal,¹⁰⁹ who, in their study of the basis set dependence of vibrational frequencies computed with the B3LYP density functional, found that “convergence of the vibrational frequencies with respect to the addition of diffuse and polarization functions is generally met already at the 6-311G(d,p) level,” based on the fact that the vibrational frequencies calculated with B3LYP/6-311G(d,p) typically agreed to within 10 cm^{-1} with those obtained from B3LYP/6-311++G(3df,3pd). All of the triple- ζ basis sets examined in the present study are larger than 6-311G(d,p), and the observed scale factor invariance, with respect to the triple- ζ basis sets used in conjunction with the various density functionals tested in this work, suggests that the convergence criteria proposed by Andersson and Uvdal in the case of B3LYP is applicable to density functionals in general.

The above observation about basis set convergence of vibrational frequencies enables us to deduce whether or not the scale factors obtained in this study, optimized based on a limited set of frequencies, reasonably represents those that would be obtained from a more extensive set because it legitimizes direct comparison with the results of other existing studies in which basis sets of at least a 6-311G(d,p) caliber were used. Andersson and Uvdal¹⁰⁹ employed 950

vibrational frequencies (belonging to 125 molecules) in their database, which led to a value for λ^{F} of 0.968 in the case of B3LYP/6-311+G(d,p). This value compares reasonably well with the λ^{F} we obtained for B3LYP/MG3S, 0.960. Another comprehensive study of scale factors has been performed by Scott and Radom¹¹ and incorporates 1066 frequencies (including degeneracies), belonging to 122 molecules, in the optimization of λ^{F} . In the tests conducted here and in the study of Scott and Radom, HF and the BLYP and B3LYP density functionals form a common subset, although different basis sets were used in the two studies. In our study, we used the MG3S basis set, while in Scott and Radom’s study, HF was paired with six different basis sets ranging from 3 to 21G to 6-311G(df,p), BLYP is included in conjunction with 6-31G(d) and 6-311G(df,p), and B3LYP is included with just the 6-31G(d) basis set. In the case of HF, we find that the values for λ^{F} of 0.9051 and 0.9054 obtained by Scott and Radom with 6-311G(d,p) and 6-311G(df,p), respectively, the only two basis sets which meet the convergence criteria set forth by Andersson and Uvdal, compare reasonably well with the value of 0.895 we obtained with the MG3S basis set. For BLYP/6-311G(df,p), Scott and Radom obtained $\lambda^{\text{F}} = 0.9986$, which is in relatively good accord with our value of 0.991. Although the 6-31G(d) basis set used with B3LYP by Scott and Radom is likely too small to have reached convergence with respect to vibrational frequencies, we note that the value of $\lambda^{\text{F}} = 0.9614$ they reported agrees well with 0.960, the value we obtained for B3LYP/MG3S. Finally, although not directly comparable, we note the similarity of our values for λ^{F} with those derived by Rauhut and Pulay for BLYP/6-31G(d) and B3LYP/6-31G(d), namely 0.995 and 0.963, respectively, for a different test set comprised of the 347 fundamental frequencies of 20 molecules.¹¹⁰ The results of these analyses indicate that the fundamental frequencies in the F38/10 database constitute a fair representative subset of a more comprehensive set, provided that one recognizes that optimal general scale factors are intrinsically uncertain by ~ 0.01 (we say “at least” because actual estimates of uncertainties (cited below) are larger). Values for λ^{H} were not derived in the studies of Andersson and Uvdal, Scott and Radom, or Rauhut and Pulay. We also note that although various integration grids have been employed in these studies, it has been shown that the effect of increasing the integration grid is usually negligible for vibrational frequencies.¹¹

Rauhut and Pulay also explored a more sophisticated scaling approach, based on the scaled quantum mechanical (SQM) force field procedure,¹¹¹ in which separate scale factors are optimized for 11 different types of internal coordinates in an effort to more closely reproduce the observed fundamentals.¹¹⁰ Although that procedure led to modest improvements, it is more complicated. To make the method more general, Rauhut and Pulay also investigated the degree of transferability of the set of internal-coordinate specific scale factors by optimizing the 11 parameters for an extended test set composed of 31 molecules, and they found that while most of these scale factors were not changed significantly, “the scaling factors for the out-of-plane modes, the torsion of conjugated systems, and the torsions of single-bonded systems show bigger deviations.”¹¹⁰ Therefore, while

Table 4. Scaled Harmonic and Fundamental Frequencies: Absolute Deviations and Outliers

model chemistry	harmonic frequencies		fundamental frequencies	
	dev ^a	N ^b	dev ^a	N ^b
B1LYP/MG3S	31 (2.2)	3	F ₂ (15), ⁴ C ₂ H ₂ (8), ² NH ₃ (-6)	F ₂ (13), Cl ₂ (-7)
B3LYP/MG3S	31 (2.1)	3	F ₂ (13), ⁴ C ₂ H ₂ (7), ² NH ₃ (-6)	F ₂ (12), Cl ₂ (-7)
B97-3/MG3S	32 (2.2)	3	F ₂ (17), ³ H ₂ CN(5), ⁴ C ₂ H ₂ (7)	F ₂ (15)
B98/MG3S	29 (1.9)	1	F ₂ (16)	F ₂ (15)
BLYP/MG3S	27 (1.4)	2	F ₂ (8), Cl ₂ (-8)	F ₂ (6), Cl ₂ (-11), ² NH ₃ (6)
BMC-CCSD	13 (0.8)	0		² NH ₃ (6)
BMK/MG3S	41 (2.7)	4	F ₂ (20), Cl ₂ (7), ⁴ C ₂ H ₂ (8), ² NH ₃ (-7)	F ₂ (19), ⁴ C ₂ H ₂ (6)
HF/MG3S	67 (4.5)	10	^{1,3} N ₂ O(6.13), F ₂ (29), N ₂ (8), ³ CO ₂ (8), ^{2,3} H ₂ CN(5.13), ^{4,5} C ₂ H ₂ (22.8), ² NH ₃ (-6)	F ₂ (27), ³ N ₂ O(9), ³ H ₂ CN(11), ⁴ C ₂ H ₂ (19), ⁵ C ₂ H ₂ (6)
HFLYP/MG3S	72 (5.0)	11	^{4,5} C ₂ H ₂ (23.7), F ₂ (31), N ₂ (8), ^{1,3} N ₂ O(7.13), ^{2,3} H ₂ CN(6.13), ³ CO ₂ (8), ² NH ₃ (-10), ² H ₂ CO(5)	F ₂ (30), N ₂ O(6), ³ N ₂ O(10), ³ H ₂ CN(11), ⁴ C ₂ H ₂ (21), ⁵ C ₂ H ₂ (6)
M05-2X/6-31+G(d,p)	42 (2.8)	5	N ₂ (6), F ₂ (24), Cl ₂ (-7), ² H ₂ O(-6), ² NH ₃ (-10)	F ₂ (22), Cl ₂ (-9), ² H ₂ O(-7)
M05-2X/def2-TZVPP	45 (3.1)	6	² NH ₃ (-8), F ₂ (26), ³ N ₂ O(7), ³ H ₂ CN(6), ⁴ C ₂ H ₂ (12), N ₂ (6)	F ₂ (24), ⁴ C ₂ H ₂ (10)
M05-2X/MG3S	43 (3.0)	6	² NH ₃ (-9), F ₂ (25), ³ N ₂ O(7), ³ H ₂ CN(6), ⁴ C ₂ H ₂ (11), N ₂ (5)	F ₂ (24), ⁴ C ₂ H ₂ (9)
M05/MG3S	47 (3.2)	7	F ₂ (17), Cl ₂ (6), ^{2,3} N ₂ O(6.9), ³ H ₂ CN(8), ⁴ C ₂ H ₂ (6), ² NH ₃ (-9)	F ₂ (16), ³ N ₂ O(6), ³ H ₂ CN(6)
M06-2X/6-31+G(d,p)	46 (3.0)	3	F ₂ (22), ⁴ C ₂ H ₂ (6), ² NH ₃ (-8)	F ₂ (21), Cl ₂ (-6), ³ CO ₂ (-6), ² H ₂ O(-6)
M06-2X/aug-cc-pVTZ	45 (3.2)	5	F ₂ (25), ³ N ₂ O(8), ³ H ₂ CN(6), ⁴ C ₂ H ₂ (12), ² NH ₃ (-6)	F ₂ (23), ⁴ C ₂ H ₂ (9)
M06-2X/def2-TZVPP	46 (3.3)	5	F ₂ (25), ³ N ₂ O(7), ³ H ₂ CN(7), ⁴ C ₂ H ₂ (13), ² NH ₃ (-6)	F ₂ (23), ⁴ C ₂ H ₂ (9)
M06-2X/maug-cc-pVTZ	45 (3.2)	4	F ₂ (25), ³ N ₂ O(7), ³ H ₂ CN(6), ⁴ C ₂ H ₂ (12), ² NH ₃ (-7)	F ₂ (23), ⁴ C ₂ H ₂ (9)
M06-2X/MG3S	45 (3.2)	6	² NH ₃ (-7), F ₂ (24), ³ N ₂ O(8), ³ H ₂ CN(7), ⁴ C ₂ H ₂ (13), N ₂ (5)	F ₂ (23), ⁴ C ₂ H ₂ (10)
M06-HF/6-31+G(d,p)	66 (4.2)	9	² NH ₃ (-11), F ₂ (31), ¹ N ₂ O(6), ⁴ C ₂ H ₂ (16), ² H ₂ O(-8), N ₂ (9), ² H ₂ CN(7), ³ H ₂ CN(7), Cl ₂ (-8)	N ₂ (6), F ₂ (29), Cl ₂ (-11), ³ CO ₂ (-7), ² H ₂ O(-8), ² N ₂ O(-8), ⁴ C ₂ H ₂ (14)
M06-HF/def2-TZVPP	65 (4.1)	8	² NH ₃ (-9), F ₂ (33), ¹ N ₂ O(6), ³ N ₂ O(8), ² H ₂ O(-6), N ₂ (8), ² H ₂ CN(6), ⁴ C ₂ H ₂ (19)	F ₂ (31), ² H ₂ O(-7), ² N ₂ O(-6), ⁴ C ₂ H ₂ (17)
M06-HF/MG3S	65 (4.1)	9	² NH ₃ (-10), F ₂ (32), N ₂ (8), ^{1,3} N ₂ O(6.8), ^{2,3} H ₂ CN(6.6), ⁴ C ₂ H ₂ (18), ² H ₂ O(-6)	F ₂ (30), Cl ₂ (-7), ² H ₂ O(-6), ² N ₂ O(-6), ⁴ C ₂ H ₂ (16)
M06-L/6-31+G(d,p)	33 (2.1)	1	⁴ C ₂ H ₂ (-7)	Cl ₂ (-6), ⁴ C ₂ H ₂ (-9)
M06-L/def2-TZVPP	35 (2.2)	5	F ₂ (7), ³ N ₂ O(7), ³ H ₂ CN(8), ⁴ C ₂ H ₂ (6), ⁵ C ₂ H ₂ (6)	³ H ₂ CN(6), ² NH ₃ (9)
M06-L/MG3S	35 (2.2)	5	F ₂ (6), ³ N ₂ O(7), ³ H ₂ CN(8), ^{4,5} C ₂ H ₂ (5.6)	³ H ₂ CN(6), ² NH ₃ (8)
M06/6-31+G(d,p)	45 (2.8)	3	F ₂ (13), ³ H ₂ CN(6), ² NH ₃ (-13)	F ₂ (12)
M06/def2-TZVPP	55 (3.6)	7	F ₂ (16), ³ N ₂ O(6), ³ N ₂ O(9), ³ H ₂ CN(10), ⁴ C ₂ H ₂ (11), ⁵ C ₂ H ₂ (6), ² NH ₃ (-8)	F ₂ (14), ³ N ₂ O(6), ³ H ₂ CN(8), ⁴ C ₂ H ₂ (9)
M06/MG3S	57 (3.6)	6	² NH ₃ (-9), F ₂ (16), ³ H ₂ CN(9), ^{2,3} N ₂ O(6.9), ⁴ C ₂ H ₂ (12)	F ₂ (14), ³ N ₂ O(6), ³ H ₂ CN(7), ⁴ C ₂ H ₂ (9)
M08-HX/6-31+G(d,p)	48 (3.0)	4	N ₂ (6), F ₂ (22), ⁴ C ₂ H ₂ (6), ² NH ₃ (-7)	F ₂ (21), Cl ₂ (-7), ³ CO ₂ (-6)
M08-HX/cc-pVTZ+	49 (3.3)	6	N ₂ (6), F ₂ (25), ¹ N ₂ O(5), ³ N ₂ O(7), ⁴ C ₂ H ₂ (11), ² NH ₃ (-6)	F ₂ (23), ³ N ₂ O(6), ⁴ C ₂ H ₂ (8)
M08-HX/def2-TZVPP	50 (3.3)	7	N ₂ (6), F ₂ (24), ¹ N ₂ O(5), ³ N ₂ O(8), ² H ₂ CN(5), ⁴ C ₂ H ₂ (10), ² NH ₃ (-5)	F ₂ (22), ³ N ₂ O(6), ⁴ C ₂ H ₂ (9)
M08-HX/MG3S	49 (3.3)	5	N ₂ (6), F ₂ (23), ³ N ₂ O(8), ⁴ C ₂ H ₂ (12), ² NH ₃ (-6)	F ₂ (22), ⁴ C ₂ H ₂ (9)
M08-SO/6-31+G(d,p)	46 (3.0)	5	F ₂ (20), ² H ₂ O(-6), ³ H ₂ CN(6), ⁴ C ₂ H ₂ (8), ² NH ₃ (-11)	F ₂ (18), ² H ₂ O(-7), ⁴ C ₂ H ₂ (6)
M08-SO/cc-pVTZ+	49 (3.4)	6	² NH ₃ (-8), F ₂ (23), ³ N ₂ O(7), ³ H ₂ CN(6), ⁴ C ₂ H ₂ (13), N ₂ (6)	F ₂ (21), ⁴ C ₂ H ₂ (11)
M08-SO/def2-TZVPP	51 (3.5)	7	² NH ₃ (-8), F ₂ (22), ¹ N ₂ O(5), ³ H ₂ CN(7), ⁴ C ₂ H ₂ (14), N ₂ (6), ³ N ₂ O(8)	F ₂ (21), ³ N ₂ O(6), ³ H ₂ CN(5), ⁴ C ₂ H ₂ (12)
M08-SO/MG3S	51 (3.5)	5	² NH ₃ (-9), F ₂ (22), ³ N ₂ O(8), ³ H ₂ CN(7), ⁴ C ₂ H ₂ (14)	F ₂ (20), ⁴ C ₂ H ₂ (12)
M08-SO/MG3SXP	53 (3.6)	6	² NH ₃ (-8), F ₂ (22), ³ N ₂ O(8), ³ H ₂ CN(7), ⁴ C ₂ H ₂ (13), N ₂ (6)	F ₂ (20), ⁴ C ₂ H ₂ (11)
PBE/MG3S	29 (1.7)	1	F ₂ (11)	F ₂ (9), Cl ₂ (-5)
PBE0/MG3S	38 (2.6)	5	² NH ₃ (-6), F ₂ (18), ³ N ₂ O(7), ³ H ₂ CN(5), ⁴ C ₂ H ₂ (7)	F ₂ (9), Cl ₂ (-5)
TPSSH/MG3S	28 (1.7)	2	F ₂ (13), ³ H ₂ CN(5)	F ₂ (16), ⁴ C ₂ H ₂ (5)
V5XC/MG3S	24 (1.6)	3	² NH ₃ (-5), F ₂ (6), Cl ₂ (-7)	F ₂ (12), Cl ₂ (-5), ² NH ₃ (6)
average	44 (2.9)	5		Cl ₂ (-10)

^a Mean absolute deviations (cm⁻¹) and, in parentheses, mean absolute % deviations after scaling. ^b N corresponds to the number of outlier frequencies, which we define as differing from experiment by more than 5%, after scaling. ^c Numbers in parentheses are the % deviations between the scaled and experimental frequencies. Superscripts are the mode numbers given in Table 1.

the SQM procedure can lead to improved accuracy, it does not do so uniformly and involves a trade-off of accuracy and simplicity.

3.1.2. Scale Factors for Zero-Point Energies. The optimal scale factors for the prediction of ZPEs from the computed harmonic vibrational frequencies obtained with the 40 electronic model chemistries tested in this study are summarized in Table 3. The rms deviation for the set is 0.41 kcal mol⁻¹, and upon scaling, this quantity is reduced to 0.18 kcal mol⁻¹. After scaling, the best electronic model chemistry tested is VSXC/MG3S, with an rms deviation of 0.08 kcal mol⁻¹. The BLYP, M06-L, and Perdew–Burke–Ernzerhof (PBE) functionals also yielded high-quality ZPEs upon scaling, with rms deviations ranging from 0.09 to 0.11 kcal mol⁻¹. The other electronic model chemistries tested also performed reasonably well on this front, with rms deviations of no more than 0.22 kcal mol⁻¹ and with the exception of those with 100% Hartree–Fock exchange (HF theory and the HFLYP and M06-HF density functionals), which yielded rms deviations that are slightly larger (0.28 – 0.30 kcal mol⁻¹). In addition, in the Supporting Information, we also report the absolute and absolute percent deviations between the experimental ZPEs and those computed with each of the 40 electronic model chemistries for all of the species studied. This analysis shows that on average these deviations amount to 0.14 ± 0.05 kcal mol⁻¹ and 2.8 ± 1.0%. Twelve of the electronic model chemistries tested yield ZPEs that, when appropriately scaled, agree with experiment to within 0.1 kcal mol⁻¹ on average. Out of these, the smallest mean absolute percent deviations result from the use of BMC–CCSD and M06-L and have a value of 1.1% in the case of the former and 1.2% in the case of the latter (when averaged over the results obtained with the three different basis sets). In particular, we identify BMC–CCSD as the most universally reliable method (of those examined here; we note that we intentionally did not examine the expensive methods that are usually affordable only for the smallest systems) for obtaining ZPEs, as evidenced by the fact that this is the only electronic model chemistry that can reproduce all of the ZPEs investigated to within 2.9%, even that of F₂, which is overestimated by the other 39 electronic model chemistries by 19.0% on average.

The λ^{ZPE} values for 17 of the 40 electronic model chemistries involved in the calculations described were also previously determined in a separate study.¹³ However, as noted earlier, these previous determinations relied on the ZPEs contained in the ZPVE15/06 database, many of which have since been revised by Irikura and co-workers.^{20,21} It is interesting to note that the previously determined values for λ^{ZPE} are only slightly different from those determined here. In fact, these differences are not just small but also systematic, with the new values being less than the old by 0.0024–0.0026 for any given electronic model chemistry, which only amounts to about a quarter of a percent difference. We further note that these two sets of scale factors are not statistically different, since the uncertainty associated with scale factors for ZPEs has been shown to be of at least ±0.02.²¹

As was the case for λ^{F} and λ^{H} , it can be seen that the values of λ^{ZPE} obtained in this study also appear to be relatively invariant with respect to the one-electron basis set used. Because of this, we can compare our scale factors to those of Scott and Radom,¹¹ as in the previous section, in order to assess if our subset of ZPEs represents a bigger data set well. In doing so, we find that the HF/MG3S λ^{ZPE} value of 0.919 obtained here agrees well with those obtained by Scott and Radom for HF/6-311G(d,p) and HF/6-311G(df,p), 0.9248 and 0.9247, respectively, and the λ^{ZPE} value of 1.013 we obtained for BLYP/MG3S is in good accord with the value of 1.0167 Scott and Radom obtained for BLYP/6-311G(df,p). In addition, we note that our λ^{ZPE} value and that of Scott and Radom for B3LYP are also in very close agreement, 0.983 and 0.9806, respectively, although Scott and Radom used a double- ζ basis set, 6-31G(d), while we employed the much larger triple- ζ basis set, MG3S. These close agreements between the λ^{ZPE} values we obtain with the 15 ZPEs in the ZPVE15/10 database and those obtained by Scott and Radom based on the larger test set of 122 molecules indicate that, in the case of ZPEs, the ZPVE15/10 database is a reasonable representative of the bigger data set.

3.1.3. Universal Scale Factor Ratios. Careful analysis of the scale factors obtained in this study lead to the realization that, for any given electronic model chemistry, knowledge of any of the three scale factors λ^{H} , λ^{F} , or λ^{ZPE} permits the accurate estimation of the other two scale factors directly, without the usual rigorous analysis of minimizing the rms deviation. Specifically, it can be seen that regardless of the electronic model chemistry used, the ratios $\lambda^{\text{H}}/\lambda^{\text{ZPE}}$ and $\lambda^{\text{F}}/\lambda^{\text{ZPE}}$ are, to a good approximation, constant, such that:

$$\lambda^{\text{H}} = \alpha^{\text{H/ZPE}} \lambda^{\text{ZPE}} \quad (13)$$

$$\lambda^{\text{F}} = \alpha^{\text{F/ZPE}} \lambda^{\text{ZPE}} \quad (14)$$

where the proportionality constants $\alpha^{\text{H/ZPE}}$ and $\alpha^{\text{F/ZPE}}$ are almost independent of the electronic model chemistry. For the set of the 40 electronic model chemistries considered in Table 3, we obtain, as shown in Table 5, the average values of 1.014 ± 0.002 and 0.974 ± 0.002 for $\alpha^{\text{H/ZPE}}$ and $\alpha^{\text{F/ZPE}}$, respectively, where the uncertainties represent 1 standard deviation (SD).

It should be emphasized that the standard deviations we attribute to these quantities only reflect the statistical scatter within a given data set, thereby neglecting any contributions from the propagation of the uncertainties in the scale factors. Irikura's statistical analyses suggest that the typical uncertainties associated with λ^{F} and λ^{ZPE} for electronic model chemistries comparable to those in this study are in the range of 0.02–0.06.^{21,112} Propagating these uncertainties indicates that our estimate of $\alpha^{\text{F/ZPE}}$ is realistically only reliable to about 3–8%. A similar measure of reliability is expected for $\alpha^{\text{H/ZPE}}$.

The value for $\alpha^{\text{F/ZPE}}$ obtained here can also be compared with those that can be derived from the studies of Pople et al.¹¹³ and Scott and Radom.¹¹ From the λ^{F} and λ^{ZPE} values obtained in the study of Pople et al., which focused on just two electronic model chemistries, and the work of Scott and

Table 5. Universal Scale Factor Ratios for Interrelating the Three Properties of Interest

model chemistry	scale factor ratios	
	$\alpha^{H/ZPE}$	$\alpha^{F/ZPE}$
B1LYP/MG3S	1.016	0.977
B3LYP/MG3S	1.016	0.977
B97-3/MG3S	1.014	0.974
B98/MG3S	1.014	0.974
BLYP/MG3S	1.017	0.979
BMC-CCSD	1.017	0.977
BMK/MG3S	1.013	0.973
HF/MG3S	1.015	0.975
HFLYP/MG3S	1.014	0.974
M05-2X/6-31+G(d,p)	1.014	0.974
M05-2X/def2-TZVPP	1.014	0.975
M05-2X/MG3S	1.013	0.974
M05/MG3S	1.013	0.973
M06-2X/6-31+G(d,p)	1.012	0.972
M06-2X/aug-cc-pVTZ	1.014	0.974
M06-2X/def2-TZVPP	1.014	0.974
M06-2X/maug-cc-pV(T+d)Z	1.013	0.974
M06-2X/MG3S	1.013	0.974
M06-HF/6-31+G(d,p)	1.015	0.976
M06-HF/def2-TZVPP	1.012	0.973
M06-HF/MG3S	1.013	0.975
M06-L/6-31+G(d,p)	1.015	0.975
M06-L/def2-TZVPP	1.020	0.980
M06-L/MG3S	1.019	0.980
M06/6-31+G(d,p)	1.009	0.970
M06/def2-TZVPP	1.014	0.974
M06/MG3S	1.014	0.974
M08-HX/6-31+G(d,p)	1.011	0.972
M08-HX/cc-pVTZ+	1.012	0.972
M08-HX/def2-TZVPP	1.011	0.972
M08-HX/MG3S	1.012	0.972
M08-SO/6-31+G(d,p)	1.010	0.971
M08-SO/cc-pVTZ+	1.013	0.973
M08-SO/def2-TZVPP	1.013	0.973
M08-SO/MG3S	1.013	0.973
M08-SO/MG3SXP	1.012	0.972
PBE/MG3S	1.015	0.976
PBE0/MG3S	1.014	0.974
TPSSH/MG3S	1.018	0.979
VSXC/MG3S	1.016	0.976
average ($\pm\sigma$) ^a	1.014(2)	0.974(2)

^a Numbers in parentheses represent the statistical uncertainties of the last significant figure.

Radom, which extended the study of Pople et al. to 17 electronic model chemistries, we derive $\alpha^{F/ZPE}$ values of 0.977 and 0.979 for their respective data sets. These values compare well with the average value of 0.974 for $\alpha^{F/ZPE}$ obtained in the present work. Values for λ^H were not reported in these^{11,113} studies, precluding the comparison of $\alpha^{H/ZPE}$ values.

It is interesting to note that our study and that of Scott and Radom differ greatly in the relative attention paid to wave function (WFT) and density functional (DFT) theories. In our study only 2 of the 40 electronic model chemistries tested were based on WFT, in particular HF and BMC-CCSD, while 11 of the 17 electronic model chemistries tested by Scott and Radom were based on WFT. Therefore, the work of Scott and Radom can be used to check whether the ratio $\alpha^{F/ZPE}$ is indeed universal or depends on whether DFT or WFT is used. We find that this ratio is virtually the same regardless of whether DFT or WFT is used, as the average

values for $\alpha^{F/ZPE}$ are 0.981 ± 0.001 and 0.978 ± 0.004 for the two respective subgroups.

3.2. Stage Two: Reduced Scale Factor Optimization

Model. As demonstrated by Irikura et al., careful consideration of the uncertainty associated with scale factors can lead to overall error margins in these quantities in the range of 0.02–0.06 for electronic model chemistries similar to the ones employed here.^{21,112} This highlights the intrinsically approximate nature of scale factors and of the resultant scaled properties, as with these kinds of error margins, many of the scale factors reported for various electronic model chemistries are in fact not statistically different. Nonetheless, it is not only customary but also very useful to have method specific scale factors optimized for the properties of interest, especially if these can be obtained at a modest computational cost.

In the preceding sections, we have shown that all three property specific scale factors can be obtained from the knowledge of just one of these by conveniently using the universal scale factor ratios established in this work. We have also shown that λ^F , λ^H , and λ^{ZPE} do not depend on the size of the test set of frequencies used in deriving them. This observation leads one to question whether the original sets of frequencies and the ZPEs can be further reduced without significantly affecting the ensuing scale factors, just as small, representative databases have been introduced in previous work^{81,114–117} for other properties.

Specifically, since λ^{ZPE} is the scale factor most commonly utilized in the literature, we investigate whether a smaller representative subset of just 6 of the 15 original ZPEs can be used to obtain values for λ^{ZPE} that agree to within 0.001 with those originally obtained for the full set of ZPEs. Out of the 5005 such possible subsets, we find that the best representative subset is that which includes the ZPEs of CH₄, NH₃, C₂H₂, H₂CO, H₂O, and N₂O, with a mean absolute deviation between λ^{ZPE} for the subset and λ^{ZPE} for the full set of 0.0007 for the 40 electronic model chemistries tested in this study. The subset of six ZPEs is called the ZPE6 database.

The finding that this subset is entirely composed of polyatomics is reasonable as one might expect that the largest ZPEs would more significantly influence the value of the scale factor by receiving the most weight in a procedure aimed at minimizing the rms deviation. However, it is interesting that the representative subset is not composed of the six molecules with the largest ZPEs but rather of five of these molecules and N₂O, which has the eighth largest ZPE. This suggests that, in general, the deviation between the computed and experimental ZPE of N₂O is larger than those for the molecules with the sixth and seventh largest ZPEs, HCN and CO₂, respectively. This hypothesis can be verified by inspecting Table S3 of the Supporting Information, from which it can be seen that the absolute deviation for the ZPE of N₂O is larger than those for the ZPEs of HCN and CO₂ for 34 of the 40 electronic model chemistries tested.

This discrepancy between the computed and experimental ZPE of N₂O might be partly due to the significant multireference character of this molecule. As has been previously noted,^{32,81} “the Hartree–Fock exchange approximation fails

Table 6. Absolute Differences between the ZPEs of Polyatomic Species Computed with BLYP/MG3S and B1LYP/MG3S^a

molecule	ZPE _B (kcal mol ⁻¹)	ZPE _{B1} (kcal mol ⁻¹)	ZPE _{B1} - ZPE _B (kcal mol ⁻¹)	100 × ZPE _{B1} - ZPE _B / ZPE _B (%)
CH ₄	27.367	28.128	0.762	2.8
NH ₃	20.918	21.607	0.689	3.3
H ₂ O	12.944	13.470	0.526	4.1
C ₂ H ₂	16.368	17.034	0.667	4.1
HCN	9.885	10.305	0.419	4.2
H ₂ CO	16.059	16.750	0.691	4.3
CO ₂	7.012	7.387	0.375	5.3
N ₂ O	6.625	7.053	0.428	6.5

^a Expressed in kcal mol⁻¹ and as a percentage.

badly for multireference systems, whereas generalized gradient approximations (GGAs) can usually handle these systems almost as well as they handle single-reference systems.” An inexpensive and useful indicator of multireference character is the B1 diagnostic,^{32,81} which relies on the comparison of the results obtained with the BLYP^{29,30} and B1LYP^{51–53} functionals, where the former is a local functional, and the latter is a hybrid functional, and the two functionals differ only in that the latter incorporates 25% of HF exchange in its design. Though the B1 diagnostic was originally defined so as to offer a semiquantitative measure of the multireference character of bond energies,^{32,81} it was later generalized to gauge the approximate extent of multireference character “for any quantity with units of energy.”¹¹⁸ Its recommended threshold of an energetic discrepancy of 10 kcal mol⁻¹ between the results of BLYP and those of B1LYP as the indicator of multireference character was, however, designed for bond energies and exceeds the values of most of the ZPEs in this study. Consequently, here we only use the concept underlying this diagnostic to qualitatively assess the degree of multireference character in the polyatomic molecules presently studied, as reflected by their ZPEs. In this analysis, we simply calculate the absolute differences between the unscaled ZPEs of the eight polyatomic species computed with BLYP/MG3S and B1LYP/MG3S. The results, presented in Table 6, show that the largest such difference between computed ZPEs does in fact occur in the case of N₂O, suggesting that out of the eight polyatomic species studied, this molecule may possess the most multireference character. The fact that CH₄ and NH₃ are at the other end of the spectrum, as might be expected, is reassuring.

Based on the generalization that the inclusion of HF exchange into electronic model chemistries can lead to bad performance when dealing with multireference systems, one would expect that HF and hybrid density functionals which incorporate the most HF exchange within their scheme will be affected the most in cases involving multireference systems. As can be seen in Table S3 of the Supporting Information, HF and HFLYP do exhibit some of the largest deviations when computing the ZPE of N₂O; however, M06-HF does not, which is consistent with the previous experience³⁴ that M06-HF is the first functional with 100% Hartree–Fock exchange that competes well with popular functionals that typically have 20–25% Hartree–Fock exchange. Furthermore, if this inability of HF exchange-based electronic model chemistries to adequately characterize multireference systems was the dominant reason for the magnitude of the deviations in the ZPE of N₂O, one would

expect M05 and M06 to outperform their 2X counterparts on this front, since 2X refers to the fact that twice the amount of HF exchange has been added in going from M0n to M0n-2X (where *n* = 5 or 6), however, this is not the case for the Minnesota functionals. Therefore, the observed deviations associated with computing the ZPE of N₂O are likely due to a combination of several factors.

Regardless of the nature of the optimal subset, the level of agreement between λ^{ZPE} for the subset and λ^{ZPE} for the full set indicates that they are often identical to the precision that we report these quantities, to the thousandths place. In addition, this signifies that no accuracy would be lost in the subsequent determinations of λ^{H} and λ^{F} through the use of universal scale factor ratios. This procedure, which we call the reduced scale factor optimization model, greatly reduces the computational cost of determining the three most important scale factors. For clarity, we summarize this reduced model as follows: (i) calculate the harmonic frequencies of CH₄, NH₃, C₂H₂, H₂CO, H₂O, and N₂O; (ii) find λ^{ZPE} by minimizing the rms deviation of eq 11 but with these 6 molecules rather than 15 (the required data are in rows 9, 10, and 12–15 of Table 2); (iii) multiply λ^{ZPE} by 1.014 and 0.974 to obtain λ^{H} and λ^{F} , respectively.

By the time this work is published, we expect to have an automated scale factor generator utility available on the Truhlar group Web site (<http://comp.chem.umn.edu/truhlar/index.htm>), which will efficiently optimize these three scale factors for any user specified electronic model chemistry through the implementation of the reduced scale factor optimization model.

3.3. Scale Factors for Additional Methods. The above analysis allows us to update Table S1 (Supporting Information) of previously computed scale factors. For any electronic model chemistry in Table S1 (Supporting Information) for which λ^{ZPE} was obtained from ZPVE15/06,¹³ we decrease λ^{ZPE} by 0.0025 (see Section 3.1.2). In addition, we have found that the scale factors obtained from the ZPVE13/99 database,^{82,119} which differs from ZPVE15/06 only in that it excludes the ZPEs of Cl₂ and OH, typically agree with those obtained from ZPVE15/06 to within 0.00001, which is well within the precision of 0.001 reported in this work. Therefore, for any electronic model chemistry in Table S1 (Supporting Information) for which λ^{ZPE} was obtained from ZPVE13/99, we also simply decrease λ^{ZPE} by 0.0025 to obtain updated quantities. For any electronic model chemistry in Table S1 (Supporting Information) for which the scale factor was not obtained from ZPVE15/06 or ZPVE13/99,

Table 7. Updated λ^{ZPE} Values to Be Used Instead of the Previous Ones Listed in Table S1 (Supporting Information)^a

model chemistry	λ^{ZPE}	% dev ^e
AM1 ^b	0.948	0.58 ^f
B1B95/6-31+G(d,p) ^c	0.971	0.26
B1B95/MG3S ^c	0.973	0.29
B3LYP/6-31(2df,2p) ^c	0.981	0.20
B3LYP/6-31G(d) ^b	0.977	0.37
B3P86/6-31G(d) ^b	0.971	0.50
B3PW91/6-31G(d) ^b	0.972	0.53
BB1K/6-31+G(d,p) ^c	0.954	0.22
BB1K/MG3S ^c	0.957	0.21
BB95/6-31+G(d,p) ^c	1.011	0.29
BB95/MG3S ^c	1.012	0.24
BLYP/6-311G(df,p) ^b	1.013	0.36
BLYP/6-31G(d) ^b	1.009	0.36
BP86/6-31G(d) ^b	1.007	0.38
HF/3-21G ^b	0.919	0.18
HF/6-31+G(d) ^b	0.911	0.58
HF/6-31+G(d,p) ^c	0.915	0.25
HF/6-311G(d,p) ^b	0.920	0.52
HF/6-311G(df,p) ^b	0.920	0.51
HF/6-31G(d) ^b	0.909	0.49
HF/6-31G(d,p) ^b	0.913	0.56
MC3BB ^c	0.965	0.26
MC3MPW ^c	0.964	0.30
MC-QCISD/3 ^c	0.992	0.20
MP2(FC)/6-31+G(d,p) ^c	0.968	0.21
MP2(FC)/6-311G(d,p) ^b	0.970	0.49
MP2(FC)/6-31G(d) ^b	0.964	0.31
MP2(FC)/6-31G(d,p) ^b	0.958	0.29
MP2(FC)/cc-pVDZ ^c	0.977	0.20
MP2(FULL)/6-31G(d) ^b	0.963	0.17
MPW1B95/6-31+G(d,p) ^c	0.970	0.22
MPW1B95/MG3S ^c	0.970	0.23
MPW1B95/MG3S ^c	0.972	0.27
MPW1K/MG3S ^c	0.953	0.23
MPW1K/MG3S ^c	0.956	0.22
MPW3LYP/6-31+G(d,p) ^c	0.980	0.25
MPW3LYP/MG3S ^c	0.982	0.26
MPWB1K/6-31+G(d,p) ^c	0.951	0.28
MPWB1K/MG3S ^c	0.954	0.28
PBE1KCIS/MG3S ^c	0.981	0.23
PBE1KCIS/MG3S ^c	0.981	0.22
PM3 ^b	0.940	3.71 ^f
PM6 ^{b,d}	1.078	9.35 ^f
PW6B95/6-31+G(d,p) ^c	0.970	0.21
QCISD(FC)/6-31G(d) ^b	0.973	0.47
X1B95/6-31+G(d,p) ^c	0.968	0.30
X1B95/MG3S ^c	0.971	0.24
XB1K/6-31+G(d,p) ^c	0.952	0.30
XB1K/MG3S ^c	0.955	0.30

^a Values for λ^{H} and λ^{F} can be obtained by multiplying λ^{ZPE} by 1.014 and 0.974, respectively, and rounding to the nearest thousandth. ^b Obtained using the reduced scale factor optimization model. ^c Obtained by decreasing the value in Table S1 (Supporting Information) by 0.0025 (see text). ^d Computed using Gaussian 09.¹²² ^e Absolute percent deviations between λ^{ZPE} values in Table 7 and Table S1 (Supporting Information). ^f Absolute percent deviations between λ^{F} values in Table 7 (0.974 \times λ^{ZPE}) and Table S1 (Supporting Information).

we compute λ^{ZPE} by steps (i) and (ii) in Section 3.2. The resulting λ^{ZPE} values are given in Table 7.

We also used the ZPE6 database to find λ^{ZPE} for electronic model chemistries not in either Tables 3 or 7. These λ^{ZPE} values are given in Table 8. Note that λ^{H} and λ^{F} for any electronic model chemistry in Tables 7 or 8 can be obtained by multiplying λ^{ZPE} by 1.014 or 0.974, respectively, and then rounding to the nearest thousandth. For example, λ^{H} and λ^{F} for MPW3LYP/6-31G(d) are 0.990 and 0.951.

Most of the scale factors in Table 7 have changed from the original scale factors in Table S1 (Supporting Information) by less than 0.6%, with the exception of PM3 and PM6,

Table 8. Additional λ^{ZPE} Values Obtained with the Reduced Scale Factor Optimization Model^a

model chemistry	λ^{ZPE}
B3LYP/ma-TZVP	0.986
B97-3/ma-TZVP ^b	0.975
B98/def2-TZVP	0.984
B98/ma-TZVP	0.985
BMK/ma-TZVP	0.972
BP86/ma-TZVP	1.014
CAM-B3LYP/ma-TZVP ^c	0.976
CCSD(T)/jul-cc-pVTZ	0.984
CCSD(T)-F12/jul-cc-pVTZ ^d	0.981
CCSD/jul-cc-pVTZ	0.973
CCSD-F12/jul-cc-pVTZ ^d	0.971
HSEh1PBE/ma-TZVP	0.979
M05/aug-cc-pVTZ	0.978
M05/ma-TZVP	0.979
M05/maug-cc-pVTZ	0.978
M05-2X/aug-cc-pVTZ	0.964
M05-2X/ma-TZVP	0.965
M05-2X/maug-cc-pVTZ	0.964
M06/aug-cc-pVTZ	0.984
M06/ma-TZVP	0.982
M06/maug-cc-pVTZ	0.982
M06-2X/ma-TZVP	0.972
M06-HF/aug-cc-pVTZ	0.961
M06-HF/ma-TZVP	0.957
M06-HF/maug-cc-pVTZ	0.959
M06-L/aug-cc-pVTZ	0.980
M06-L/ma-TZVP	0.977
M06-L/maug-cc-pVTZ	0.977
M08-HX/aug-cc-pVTZ	0.975
M08-HX/ma-TZVP	0.976
M08-HX/maug-cc-pVTZ	0.976
M08-SO/aug-cc-pVTZ	0.985
M08-SO/ma-TZVP	0.984
M08-SO/maug-cc-pVTZ	0.983
MOHLYP/ma-TZVP	1.027
MOHLYP/MG3S	1.022
MP4(SDQ)/jul-cc-pVTZ	0.973
MPW1K/ma-TZVP	0.956
MPW1K/MIDI!	0.953
MPW1K/MIDIY	0.947
MPW3LYP/6-311+G(2d,p)	0.986
MPW3LYP/6-31G(d)	0.976
MPW3LYP/ma-TZVP	0.986
MPWLYP1M/ma-TZVP	1.009
SOGGA/ma-TZVP	1.017
τ -HCTHhyb/ma-TZVP	0.989
TPSS1KCIS/def2-TZVP	0.982
TPSS1KCIS/ma-TZVP	0.983
ω B97/def2-TZVP ^c	0.969
ω B97/ma-TZVP ^c	0.970
ω B97X/def2-TZVP ^c	0.970
ω B97X/ma-TZVP ^c	0.971
ω B97X-D/def2-TZVP ^c	0.975
ω B97X-D/ma-TZVP ^c	0.975
ω B97X-D/maug-cc-pVTZ ^c	0.974

^a Values for λ^{H} and λ^{F} can be obtained by multiplying λ^{ZPE} by 1.014 and 0.974, respectively, and rounding to the nearest thousandth. ^b Computed using locally modified version of Gaussian 03.^{106,123} ^c Computed using Gaussian 09.¹²² ^d Computed using MOLPRO 2009.1.¹²⁴

where the percentage changes are 4 and 9%, respectively. For methods based on the neglect-of-diatomic differential overlap, the percentage changes we report in Table 7 are actually for λ^{F} (obtained by multiplying λ^{ZPE} by 0.974), since this is the quantity that is reported for these methods in Table S1 (Supporting Information). The λ^{F} values in Table S1 (Supporting Information) come from the work of Scott and Radom¹¹ for AM1 and PM3 and from the work of Fekete et al.¹²⁰ for PM6 and were directly obtained from rms deviation minimization procedures, rather than the indirect way in which we obtain them here via the reduced scale factor optimization model (steps i–iii of Section 3.2). Although the difference between the λ^{F} values in Tables 7 and S1 (Supporting Information) for AM1 is reasonable, 0.6%, the

Table 9. Comparison between Experimental and Calculated ZPEs Used in Deriving Scale Factors and Universal Scale Factor Ratios in the Present Study

species	ZPE _(exp) ^b (cm ⁻¹)	SD (1σ) (cm ⁻¹)	ZPE _(calc) (cm ⁻¹)	absolute deviation ^a		
				cm ⁻¹	kcal mol ⁻¹	%
CO ₂	2554	80	2575	21	0.06	0.81
H ₂ O	4636	10	4624	12	0.03	0.25
N ₂ O	2368	N/A	2406	38	0.11	1.62
HCN	3508	111	3504	4	0.01	0.13
C ₂ H ₂	5768	2	5812	43	0.12	0.75
NH ₃	7415	N/A	7407	8	0.02	0.10
CH ₄	9692	N/A	9733	41	0.12	0.43
average				24	0.07	0.58

^a These values represent the absolute deviations from the central value from the range of the experimental ZPE ± 1σ. ^b See text for references.

large discrepancies between the λ^F values obtained for PM3 and PM6 from the reduced scale factor optimization model and the original direct rms deviation minimization procedures suggest that errors in the PM3 and PM6 methods are less systematic than those in the other methods studied here. In addition, we note that even upon scaling, neglect-of-diatomic differential-overlap methods give very unreliable vibrational frequencies, with rms deviations of 126, 159, and 96 cm⁻¹ for AM1, PM3, and PM6, respectively,^{11,120} and such methods should be used for computational thermochemistry with extreme caution or even avoided.

4. Applications of Universal Scale Factor Ratios

The universal scale factor ratios can have a wide range of applications. In addition to providing a much more convenient and inexpensive method for obtaining all three scale factors discussed in this paper for any given electronic model chemistry without appreciably compromising the accuracy, the universal scale factor ratios can also be used to estimate the ZPE of any molecule directly from its observed fundamental and/or true harmonic vibrational frequencies:

$$\varepsilon_{\text{vib}}^{\text{G}} \cong \frac{\hbar c}{2\alpha^{\text{F/ZPE}}} \sum_m v_m \quad (15)$$

$$\varepsilon_{\text{vib}}^{\text{G}} \cong \frac{\hbar c}{2\alpha^{\text{H/ZPE}}} \sum_m \omega_m \quad (16)$$

The above equations can be useful for a wide range of scenarios, but eq 15 is especially valuable for approximating the ZPE of polyatomic species whose fundamental frequencies are known, but the ZPEs of which may not be available due to insufficient spectroscopic data. In turn, knowledge of the ZPE can be useful for a number of applications. For instance, to more accurately gauge the performance of a given electronic model chemistry with regards to heats of reactions, subtracting the experimental ZPEs from the θ K heats of reaction of the products and reactants will allow for the direct comparison of the resulting electronic energies.

As a practical example, we now aim to quantify the degree of accuracy that can be achieved for predicting the ZPEs of polyatomic molecules through the use of α^{F/ZPE} (i.e., via eq 15). We do this through two separate analyses. In the first analysis, we directly compare the ZPEs calculated via eq 15

Table 10. Comparison between Experimental and Calculated ZPEs Not Used in Deriving Scale Factors and Universal Scale Factor Ratios in the Present Study

species	ZPE _(exp) ^b (cm ⁻¹)	SD (1σ) (cm ⁻¹)	ZPE _(calc) (cm ⁻¹)	absolute deviation ^a		
				cm ⁻¹	kcal mol ⁻¹	%
OCS	2016	63	2033	17	0.05	0.86
SO ₂	1542	48	1556	14	0.04	0.88
H ₂ S	3316	107	3298	18	0.05	0.54
CS ₂	1520	48	1533	14	0.04	0.91
NO ₂	1889	62	1892	3	0.01	0.15
CICN	1888	59	1908	20	0.06	1.06
HOCl	2863	9	2860	3	0.01	0.10
HOF	3026	96	2986	40	0.11	1.32
average				16	0.05	0.73

^a These values represent the absolute deviations from the central value of the range from the experimental ZPE ± 1σ. ^b See text for references.

with their experimental counterparts for the polyatomic species in the ZPVE15/10 database, with the exception of H₂CO, the experimental ZPE of which has an uncertainty that is too large for a meaningful comparison. In the second analysis of this process, we compare the calculated and experimental ZPEs of eight polyatomic species that were not used in the scale factor optimization or universal scale factor ratio analyses in this work, namely OCS, SO₂, H₂S, CS₂, NO₂, CICN, HOCl, and HOF. The experimental ZPEs of these species constitute the only remaining available new estimates from Irikura et al.,²¹ with the exception of C₂H₄ and CH₃Cl, whose experimental ZPEs we omit because they have very large error bars. The accuracy with which the ZPEs of these eight species can be predicted, via eq 15, should be a reasonable indicator of the general reliability and transferability of α^{F/ZPE}. All experimentally observed fundamental frequencies needed for these two analyses were taken from the Computational Chemistry Comparison and Benchmark Database,¹⁵ and the ensuing results are given in Tables 9 and 10. The excellent agreement between the calculated and experimental ZPEs, in both cases of ZPEs, which were used to establish scale factors and universal scale factor ratios, mean absolute deviation of 0.07 kcal mol⁻¹ (0.58%), and of the ZPEs that were not used to establish these quantities, mean absolute deviation of 0.05 kcal mol⁻¹ (0.73%), indicates that α^{F/ZPE} is reasonably transferable and reliable for predicting accurate ZPEs for stable covalently bound polyatomic species from their observed fundamental frequen-

cies, provided that these frequencies are also reliable. Averaging over the results in both tables (i.e., all 15 polyatomic ZPEs) leads to the value of 0.06 kcal mol⁻¹ (0.66%) for the mean absolute deviation and 0.07 kcal mol⁻¹ (0.80%) for the rms deviation. We suggest taking twice the rms deviation as the 95% confidence limits ($2\sigma = 1.6\%$) when using eq 15 to approximate experimental ZPEs.

5. Concluding Remarks

Although we encourage the use of these newly optimized scale factors and their ratios, we point out that while these can usually lead to accurate predictions of the quantities of interest, for some specific problems, the use of a general scale factor can result in large discrepancies between the computed and experimental properties of interest. For example, some of the electronic model chemistries tested in this work overestimated the harmonic and/or fundamental stretching frequency in F₂ by more than 20% and its ZPE by more than 0.3 kcal mol⁻¹ (23%). An error of this magnitude can propagate and lead to a significant contribution to the overall error of a thermochemical quantity of interest. It is important to keep in mind that throughout this study, the minimizations of the rms deviations are done with respect to absolute errors, which are generally the greatest for the largest members in the set. Therefore, because the values of the vibrational frequency and ZPE in F₂ are quite small, even a 20% error in either of these quantities does not amount to as much as just a 5% error in the frequency of H₂ or a 1% error in the ZPE of CH₄. An alternative would be to minimize the rms deviations of a set of relative errors, in which case the large fractional deviations in F₂ discussed above would receive more weight than the smaller percentage errors in the equivalent properties of other molecules.

It should also be emphasized that because vibrational frequencies and ZPEs of weakly bound and/or noncovalent species are generally not available, these types of species have been excluded from the present study. Due to the loose vibrations possessed by these species, the scale factors and scale factor ratios established in this work based on species predominantly characterized by strong covalent interactions are probably inapplicable to thermochemical calculations dominated by the low frequencies of noncovalently bonded species.

Finally, it is well recognized that for severely anharmonic frequencies, simply scaling frequencies computed within the harmonic approximation can often be inadequate. These types of frequencies include many torsions and most inversions, such as the umbrella modes of halomethyl radicals.¹²¹ More sophisticated scaling methods which treat these kinds of modes separately in the scaling procedure or, indeed, which treat several different classes of internal coordinates separately can improve upon the accuracy that can be achieved with just one general scale factor, though this improvement is often just marginal in the case of fundamental frequencies, and as has been noted, “for the calculation of zero-point energies and thermodynamic parameters, single scale factors are adequate.”¹¹⁰ However, in principle, precautionary diagnostic analyses should always be conducted beforehand, when feasible, to determine the degree of applicability of these scale factors and universal ratios to the systems of interest.

6. Summary

The optimal scale factors for reproducing vibrational harmonic frequencies, fundamental frequencies, and zero-point energies (ZPEs) from computed harmonic frequencies have been obtained for 145 electronic model chemistries, where an electronic model chemistry is a combination of an electronic structure level or density functional approximation with a one-electron basis or is a multicoefficient correlation method, a doubly hybrid density functional, or a neglect-of-diatomic differential-overlap model. Extensive statistics for large databases were used for 40 of these electronic model chemistries. In these cases, the experimental values for true harmonic frequencies can be reproduced to typically within 2.9%, fundamental frequencies can be predicted typically to within 2.4%, and the ZPE can be estimated to typically within 0.14 kcal mol⁻¹ from the appropriately scaled computed harmonic frequencies. The frequencies computed with the balanced multicoefficient correlation method based on coupled cluster theory with single and double excitations (BMC-CCSD) multilevel method, scaled accordingly, were found to give the most reliable results overall. The M06-L and VSXC functionals also consistently yielded high-quality results.

The scale factor ratios λ^H/λ^{ZPE} and λ^F/λ^{ZPE} were found to be invariant to a good approximation with respect to the electronic model chemistry used to compute the scale factors. In light of this, we denote these universal quantities as $\alpha^{H/ZPE}$ and $\alpha^{F/ZPE}$, and we recommend using their respective average values of 1.014 and 0.974. These values can be used to interconvert between scale factors optimized for various properties, thereby allowing the extraction of all three scale factors for the cost of just one. Additional computational cost can be saved when the λ^{ZPE} used to obtain the other scale factors is obtained from the small representative database (called ZPE6) presented here.

Acknowledgment. This material is based on work supported as part of the Combustion Energy Frontier Research Center, funded by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences under Award No. DE-SC 0001198.

Supporting Information Available: Previous scale factors and absolute ZPE deviations for new scale factors. This information is available free of charge via the Internet at <http://pubs.acs.org/>.

References

- (1) Hehre, W. J.; Radom, L.; Schleyer, P. v. R.; Pople, J. A. *Ab Initio Molecular Orbital Theory*; Wiley: New York, 1986.
- (2) Zhang, Q.; Day, P. N.; Truhlar, D. G. *J. Chem. Phys.* **1993**, *98*, 4948.
- (3) Kuhler, K. M.; Truhlar, D. G.; Isaacson, A. D. *J. Chem. Phys.* **1996**, *104*, 4664.
- (4) Isaacson, A. D. *J. Chem. Phys.* **1998**, *108*, 9978.
- (5) Tew, D. P.; Handy, N. C.; Carter, S. *Phys. Chem. Chem. Phys.* **2001**, *3*, 1958.
- (6) Barone, V. *J. Chem. Phys.* **2005**, *122*, 014108.
- (7) Pople, J. A. *Rev. Mol. Phys.* **1999**, *71*, 267.

- (8) Lynch, B. J.; Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2005**, *109*, 1643.
- (9) Zhao, Y.; Lynch, B. J.; Truhlar, D. G. *J. Phys. Chem. A* **2004**, *108*, 4786.
- (10) Grev, R. S.; Janssen, C. L.; Schaefer, H. F., III. *J. Chem. Phys.* **1991**, *95*, 5128.
- (11) Scott, A. P.; Radom, L. *J. Phys. Chem.* **1996**, *100*, 16502.
- (12) Herzberg, G. *Infrared and Raman Spectra of Polyatomic Molecules*; Van Nostrand: Princeton, NJ, 1945.
- (13) Zhao, Y.; Truhlar, D. G. *Theor. Chem. Acc.* **2008**, *120*, 215.
- (14) Martin, J. M. L.; El-Yazal, J.; Francois, J.-P. *Mol. Phys.* **1995**, *86*, 1437.
- (15) Johnson, R. D., III Computational Chemistry Comparison and Benchmark Database, version 14; National Institute of Standards and Technology: Gaithersburg, MD; <http://cccbdb.nist.gov>. Accessed February 9, 2010.
- (16) *NIST Chemistry Web Book*; National Institute of Standards and Technology: Gaithersburg, MD; <http://webbook.nist.gov/chemistry>. Accessed April 21, 2010.
- (17) Mead, C. A. In *Mathematical Frontiers in Computational Chemical Physics*; Truhlar, D. G., Ed.; Springer-Verlag: New York, 1988; IMA Vol. 15, pp 1–17.
- (18) Ballhausen, C. J.; Hansen, A. E. *Annu. Rev. Phys. Chem.* **1972**, *23*, 15.
- (19) Martin, J. M. L. *J. Chem. Phys.* **1992**, *97*, 5012.
- (20) Irikura, K. K. *J. Phys. Chem. Ref. Data* **2007**, *36*, 389.
- (21) Irikura, K. K.; Johnson, R. D., III; Kacker, R. N.; Kessel, R. *J. Chem. Phys.* **2009**, *130*, 114102.
- (22) Truhlar, D. G.; Isaacson, A. D. *J. Chem. Phys.* **1991**, *94*, 357.
- (23) Barone, V. *J. Chem. Phys.* **2004**, *120*, 3059.
- (24) Schuurman, M. S.; Muir, S. R.; Allen, W. D.; Schaefer, H. F., III. *J. Chem. Phys.* **2004**, *120*, 11586.
- (25) Tajti, A.; Szalay, P. G.; Csaszar, A. G.; Kallay, M.; Gauss, J.; Valeev, E. F.; Flowers, B. A. J., V.; Stanton, J. F. *J. Chem. Phys.* **2004**, *121*, 11599.
- (26) Csonka, G. I.; Ruzsinszky, A.; Perdew, J. P. *J. Phys. Chem. A* **2005**, *109*, 6779.
- (27) Huber, K. P.; Herzberg, G. *Molecular Spectra and Molecular Structure: IV. Constants of Diatomic Molecules*; van Nostrand Reinhold: New York, 1979.
- (28) Dunham, J. L. *Phys. Rev.* **1932**, *41*, 721.
- (29) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1988**, *37*, 785.
- (30) Becke, A. D. *Phys. Rev. A: At., Mol., Opt. Phys.* **1988**, *38*, 3098.
- (31) Adamo, C.; Barone, V. *Chem. Phys. Lett.* **1997**, *274*, 242.
- (32) Zhao, Y.; Schultz, N. E.; Truhlar, D. G. *J. Chem. Theory Comput.* **2006**, *2*, 364.
- (33) Zhao, Y.; Truhlar, D. G. *J. Chem. Phys.* **2006**, *125*, 194101.
- (34) Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2006**, *110*, 13126.
- (35) Zhao, Y.; Truhlar, D. G. *J. Chem. Theory Comput.* **2008**, *4*, 1849.
- (36) Hehre, W. J.; Ditchfield, R.; Pople, J. A. *J. Chem. Phys.* **1972**, *56*, 2257.
- (37) Lynch, B. J.; Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2003**, *107*, 1384.
- (38) Weigend, F.; Haser, M.; Patzelt, H.; Ahlrichs, R. *Chem. Phys. Lett.* **1998**, *294*, 143.
- (39) Weigend, F.; Ahlrichs, R. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3297.
- (40) Dunning, T. H., Jr. *J. Chem. Phys.* **1989**, *90*, 1007.
- (41) Clark, T.; Chandrasekhar, J.; Spitznagel, G. W.; Schleyer, P. v. R. *J. Comput. Chem.* **1983**, *4*, 294.
- (42) Woon, D. E.; Dunning, T. H., Jr. *J. Chem. Phys.* **1993**, *98*, 1358.
- (43) Papajak, E.; Leverentz, H. R.; Zheng, J.; Truhlar, D. G. *J. Chem. Theory Comput.* **2009**, *5*, 1197.
- (44) Kendall, R. A.; Dunning, T. H., Jr.; Harrison, R. J. *J. Chem. Phys.* **1992**, *96*, 6796.
- (45) Dunning, T. H., Jr.; Peterson, K. A.; Wilson, A. K. *J. Chem. Phys.* **2001**, *114*, 9244.
- (46) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648.
- (47) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. *J. Phys. Chem.* **1994**, *98*, 11623.
- (48) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865.
- (49) Schmider, H. L.; Becke, A. D. *J. Chem. Phys.* **1998**, *108*, 9624.
- (50) Voorhis, T. V.; Scuseria, G. E. *J. Chem. Phys.* **1998**, *109*, 400.
- (51) Adamo, C.; Barone, V. *J. Chem. Phys.* **1999**, *110*, 6158.
- (52) Adamo, C.; Cossi, M.; Barone, V. *THEOCHEM* **1999**, *493*, 147.
- (53) Ernzerhof, M.; Scuseria, G. E. *J. Chem. Phys.* **1999**, *110*, 5029.
- (54) Barone, V.; Fliszar, S. *THEOCHEM* **1996**, *369*, 29.
- (55) Valentin, C. D.; Pacchioni, G.; Bredow, T.; Dominguez-Ariza, D.; Illas, F. *J. Chem. Phys.* **2002**, *117*, 2299.
- (56) Staroverov, V. N.; Scuseria, G. E.; Tao, J.; Perdew, J. P. *J. Chem. Phys.* **2003**, *119*, 12129.
- (57) Boese, A. D.; Martin, J. M. L. *J. Chem. Phys.* **2004**, *121*, 3405.
- (58) Keal, T. W.; Tozer, D. J. *J. Chem. Phys.* **2005**, *123*, 121103.
- (59) Zhao, Y.; Schultz, N. E.; Truhlar, D. G. *J. Chem. Phys.* **2005**, *123*, 161103.
- (60) Pople, J. A.; Head-Gordon, M.; Raghavachari, K. *J. Chem. Phys.* **1987**, *87*, 5968.
- (61) Purvis, G. D.; Bartlett, R. J. *J. Chem. Phys.* **1982**, *76*, 1910.
- (62) Raghavachari, K.; Trucks, G. W.; Pople, J. A.; Head-Gordon, M. *Chem. Phys. Lett.* **1989**, *157*, 479.
- (63) Adler, T. B.; Knizia, G.; Werner, H.-J. *J. Chem. Phys.* **2007**, *127*, 221106.
- (64) Knizia, G.; Adler, T. B.; Werner, H.-J. *J. Chem. Phys.* **2009**, *130*, 054104.
- (65) Becke, A. D. *J. Chem. Phys.* **1996**, *104*, 1040.
- (66) Perdew, J. P. *Phys. Rev. B* **1986**, *33*, 8822.
- (67) Perdew, J. P. In *Electronic Structure of Solids '91*; Ziesche, P., Eschrig, H., Eds.; Akademie Verlag: Berlin, 1991; pp 11.

- (68) Perdew, J. P.; Chevary, J. A.; Vosko, S. H.; Jackson, K. A.; Pederson, M. R.; Singh, D. J.; Fiolhais, C. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1992**, *46*, 6671.
- (69) Perdew, J. P.; Chevary, J. A.; Vosko, S. H.; Jackson, K. A.; Pederson, M. R.; Singh, D. J.; Fiolhais, C. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1993**, *48*, 4978.
- (70) Perdew, J. P.; Burke, K.; Wang, Y. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1996**, *54*, 16533.
- (71) Burke, K.; Perdew, J. P.; Wang, Y. In *Electronic Density Functional Theory: Recent Progress and New Directions*; Dobson, J. F., Vignale, G., Das, M. P., Eds.; Plenum: New York, 1998.
- (72) Zhao, Y.; Lynch, B. J.; Truhlar, D. G. *J. Phys. Chem. A* **2004**, *108*, 2715.
- (73) Yanai, T.; Tew, D.; Handy, N. *Chem. Phys. Lett.* **2004**, 393, 51.
- (74) Heyd, J.; Scuseria, G. E. *J. Chem. Phys.* **2004**, *121*, 1187.
- (75) Heyd, J.; Scuseria, G. E. *J. Chem. Phys.* **2004**, *120*, 7274.
- (76) Heyd, J.; Peralta, J. E.; Scuseria, G. E.; Martin, R. L. *J. Chem. Phys.* **2005**, *123*, 174101: 1.
- (77) Heyd, J.; Scuseria, G. E.; Ernzerhof, M. *J. Chem. Phys.* **2006**, *124*, 219906.
- (78) Izmaylov, A. F.; Scuseria, G. E.; Frisch, M. J. *J. Chem. Phys.* **2006**, *125*, 104103: 1.
- (79) Krukau, A. V.; Vydrov, O. A.; Izmaylov, A. F.; Scuseria, G. E. *J. Chem. Phys.* **2006**, *125*, 224106.
- (80) Henderson, T. M.; Izmaylov, A. F.; Scalmani, G.; Scuseria, G. E. *J. Chem. Phys.* **2009**, *131*, 044108.
- (81) Schultz, N. E.; Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2005**, *109*, 11127.
- (82) Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2004**, *108*, 6908.
- (83) Adamo, C.; Barone, V. *J. Chem. Phys.* **1998**, *108*, 664.
- (84) Boese, A. D.; Handy, N. C. *J. Chem. Phys.* **2002**, *116*, 9559.
- (85) Lynch, B. J.; Fast, P. L.; Harris, M.; Truhlar, D. G. *J. Phys. Chem. A* **2000**, *104*, 4811.
- (86) Xu, X.; Goddard III, W. A. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 2673.
- (87) Krieger, J. B.; Chen, J.; Iafate, G. J.; Savin, A. In *Electron Correlations and Materials Properties*; Gonis, A., Kiousis, N., Eds.; Plenum: New York, 1999; pp 463.
- (88) Zhao, Y.; Truhlar, D. G. *J. Chem. Theory Comput.* **2005**, *1*, 415.
- (89) Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2005**, *109*, 5656.
- (90) Zhao, Y.; Truhlar, D. G. *J. Chem. Phys.* **2008**, *128*, 184109.
- (91) Tao, J.; Perdew, J. P.; Staroverov, V. N.; Scuseria, G. E. *Phys. Rev. Lett.* **2003**, *91*, 146401.
- (92) Zhao, Y.; Lynch, B. J.; Truhlar, D. G. *Phys. Chem. Chem. Phys.* **2005**, *7*, 43.
- (93) Chai, J.-D.; Head-Gordon, M. *J. Chem. Phys.* **2008**, *128*, 084106.
- (94) Chai, J.-D.; Head-Gordon, M. *Phys. Chem. Chem. Phys.* **2008**, *10*, 6615.
- (95) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F. *J. Am. Chem. Soc.* **1985**, *107*, 3902.
- (96) Stewart, J. J. P. *J. Comput. Chem.* **1989**, *10*, 209.
- (97) Stewart, J. J. P. *J. Comput. Chem.* **1989**, *10*, 221.
- (98) Stewart, J. J. P. *J. Mol. Model.* **2007**, *13*, 1173.
- (99) Lynch, B. J.; Truhlar, D. G. *J. Phys. Chem. A* **2003**, *107*, 3898.
- (100) Zheng, J.; Xu, X.; Truhlar, D. G. *Theor. Chem. Acc.*, to be submitted
- (101) Papajak, E.; Truhlar, D. G. *J. Chem. Theory Comput.* **2010**, *6*, 597.
- (102) Tatewaki, H.; Huzinaga, S. *J. Comput. Chem.* **1980**, *1*, 205.
- (103) Easton, R. E.; Giesen, D. J.; Welch, A.; Cramer, C. J.; Truhlar, D. G. *Theor. Chim. Acta.* **1996**, *93*, 281.
- (104) Lynch, B. J.; Truhlar, D. G. *Theor. Chem. Acc.* **2004**, *111*, 335.
- (105) Zheng, J.; Alecu, I. M.; Lynch, B. J.; Zhao, Y.; Truhlar, D. G. Database of Frequency Scaling Factors for Electronic Structure Methods, version 1; University of Minnesota: Minneapolis, MN; http://comp.chem.umn.edu/truhlar/freq_scale.htm. Accessed November 15, 2009.
- (106) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A. 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. 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.; A. A.-L. M.; 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.
- (107) Zhao, Y.; Truhlar, D. G. *MN-GFM: Minnesota Gaussian Functional Module*, version 4.1; University of Minnesota: Minneapolis, MN, 2008.
- (108) Zhao, Y.; Truhlar, D. G. *MLGAUSS*, version 2.0; University of Minnesota: Minneapolis, MN, 2005.
- (109) Andersson, M. P.; Uvdal, P. *J. Phys. Chem. A* **2005**, *109*, 2937.
- (110) Rauhut, G.; Pulay, P. *J. Phys. Chem.* **1995**, *99*, 3093.
- (111) Pulay, P.; Fogarasi, G.; Pongor, G.; Boggs, J. E.; Vargha, A. *J. Am. Chem. Soc.* **1983**, *105*, 7073.
- (112) Irikura, K. K.; Johnson, R. D., III; Kacker, R. N. *J. Phys. Chem. A* **2005**, *109*, 8430.
- (113) Pople, J. A.; Scott, A. P.; Wong, M. W.; Radom, L. *Isr. J. Chem.* **1993**, *33*, 345.
- (114) Lynch, B. J.; Truhlar, D. G. *J. Phys. Chem. A* **2003**, *107*, 8996.
- (115) Schultz, N. E.; Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2005**, *109*, 4388.
- (116) Zheng, J.; Zhao, Y.; Truhlar, D. G. *J. Chem. Theory Comput.* **2007**, *3*, 569.

- (117) Zheng, J.; Zhao, Y.; Truhlar, D. G. *J. Chem. Theory Comput.* **2009**, *5*, 808.
- (118) Zhao, Y.; Tishchenko, O.; Gour, J. R.; Li, W.; Lutz, J. J.; Piecuch, P.; Truhlar, D. G. *J. Phys. Chem. A* **2009**, *113*, 5786.
- (119) Fast, P. L.; Corchado, J.; Sanchez, M. L.; Truhlar, D. G. *J. Phys. Chem. A* **1999**, *103*, 3139.
- (120) Fekete, Z. A.; Hoffmann, E. A.; Kortvelyesi, T.; Penke, B. *Mol. Phys.* **2007**, *105*, 2597.
- (121) Marshall, P.; Srinivas, G. N.; Schwartz, M. *J. Phys. Chem. A* **2005**, *109*, 6371.
- (122) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A. J.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, O.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. *Gaussian 09*, revision A.02; Gaussian, Inc.: Wallingford, CT, 2009.
- (123) Zhao, Y.; Truhlar, D. G. *MN-GFM: Minnesota Gaussian Functional Module*, version 4.3; University of Minnesota: Minneapolis, MN, 2009.
- (124) Werner, H.-J.; Knowles, P. J.; Lindh, R.; Manby, F. R.; Schutz, M.; Celani, P.; Korona, T.; Mitrushenkov, A.; Rauhut, G.; Adler, T. B.; Amos, R. D.; Bernhardsson, A.; Berning, A.; Cooper, D. L.; Deegan, M. J. O.; Dobbyn, A. J.; Eckert, F.; Goll, E.; Hampel, C.; Hetzer, G.; Hrenar, T.; Knizia, G.; Koppl, C.; Liu, Y.; Lloyd, A. W.; Mata, R. A.; May, A. J.; McNicholas, S. J.; Meyer, W.; Mura, M. E.; Nicklass, A.; Palmieri, P.; Pfluger, K.; Pitzer, R.; Reiher, M.; Schumann, U.; Stoll, H.; Stone, A. J.; Tarroni, R.; Thorsteinsson, T.; Wang, M.; Wolf, A. *MOLPRO*, version 2009.1; Cardiff University and University of Stuttgart: Wales, U.K. and Stuttgart, Germany package of ab initio programs, 2009.

CT100326H

Ab Initio Study of Water Polarization in the Hydration Shell of Aqueous Hydroxide: Comparison between Polarizable and Nonpolarizable Water Models

Denis Bucher,[†] Angus Gray-Weale,[‡] and Serdar Kuyucak*

School of Physics, University of Sydney, NSW 2006, Australia

Received July 1, 2010

Abstract: Ab initio simulations of aqueous hydroxide are performed to study the structure and polarization of water molecules in the first solvation shell. Polarization is found to depend on the configuration of the hydrogen-bond (HB) donors. In the most common case of four HB donors, the dipole moment of water molecules is much larger than those in the first shell of monovalent ions. When there are only three HB donors, the water dipole moment exceeds even those in the first shell of a divalent cation. We also show that the dipole fluctuations in the first hydration shell of hydroxide are reduced compared to bulk water, which can provide a rationale for the propensity of hydroxide for interfaces with hydrophobes. Because of its unique properties, hydroxide provides a nontrivial test for benchmarking classical models. Comparison of the ab initio results with those obtained from the classical models indicates that the latter need to be further improved in order to yield reliable results.

Introduction

Water's autolysis into hydronium (H_3O^+) and hydroxide (OH^-) ions is important because these ions affect and even control many biological and chemical processes. While the fast diffusion of these ions relative to water has been explained via the Grotthuss mechanism,¹ involving transfer of a proton along a hydrogen bond (HB), a truly microscopic understanding of this mechanism has only been obtained recently with the application of the ab initio molecular dynamics (AIMD) simulations to aqueous solutions. Proton solvation and transport, in particular, has attracted a great deal of attention and has been studied by both ab initio^{2–7} and empirical valence bond models.^{8–15} According to the picture emerging from these studies, there is no dominant structure for the hydronium ion, but rather it continuously evolves between the Eigen complex with three HB acceptors, $[\text{H}_3\text{O}^+(\text{H}_2\text{O})_3]$ and the Zundel cation, $[\text{H}_2\text{O}\cdots\text{H}\cdots\text{OH}_2]$. Proton transfer from the hydronium ion to a neighboring

water molecule is facilitated when one of the HB's in the receiving water breaks so that it attains a similar solvation structure to that of hydronium (according to the presolvation concept).¹⁶

From symmetry arguments, a similar mechanism involving proton holes had been assumed to operate in hydroxide transport.^{17,18} In this mirror-image analogy, hydroxide accepts three HB's and donates none, and a proton is transferred from a neighboring water molecule to the hydroxide along the HB. A difficulty with this mechanism is that, while the donor water is tetrahedrally coordinated, the water derived from the hydroxide after the proton transfer is only three-fold coordinated, in conflict with the presolvation concept. Indeed, the pioneering AIMD simulations of aqueous hydroxide indicated that the situation is much more complex than that suggested by the mirror mechanism.^{2,16,19,20} The dominant solvation configuration is hypercoordinated with four HB donors, and the three-fold coordination occurs intermittently as a result of fluctuations. Also the hydroxide hydrogen is a HB donor most of the time. The difficulty of obtaining the right solvation structure for hydroxide was highlighted when another AIMD simulation of hydroxide using a different density functional (PW91²¹ instead of BLYP)^{22,23} found the three-fold coordination as the dominant

* Corresponding author. E-mail: serdar@physics.usyd.edu.au. Telephone: (61) 2 9036 5306.

[†] Present address: Department of Chemistry and Biochemistry, University of California, San Diego, La Jolla, CA 92093-0365.

[‡] Present address: School of Chemistry, Monash University, Clayton 3800, Victoria, Australia.

structure.²⁴ Further investigation of this issue including a third density functional (HCTH)²⁵ confirmed the previous solvation structures obtained with BLYP and PW91 and further showed that use of the HCTH functional led to a further stabilization of the configuration with four HB donors.²⁶ The question of which functional to use was settled by looking at the dynamics—PW91 led to an ultrafast diffusion of hydroxide, while that of HCTH was too slow, and only BLYP gave a diffusion coefficient in reasonable agreement with the experimental data.^{26,27} This suggests that the solvation structure obtained using BLYP is the most realistic one. This structure is also supported by the neutron diffraction^{28,29} and solution X-ray diffraction experiments,³⁰ which are consistent with four strongly bound HB donors and a more weakly bound HB acceptor.

A recent debate about the surface affinity of hydronium and hydroxide ions has accentuated the need for accurate modeling of these ions. From oil droplet and air bubble experiments,^{31,32} it has long been known that the surface of neat water is negatively charged, indicating that hydroxide has a higher affinity for a hydrophobic surface than hydronium. In contrast, recent spectroscopic experiments based on second harmonic generation have found that hydronium is enhanced at the surface, while hydroxide is not.^{33,34} More recent spectroscopic work did detect a strong hydroxide adsorption at an octadecyltrichlorosilane–water interface,³⁵ in agreement with predictions of electrokinetic and stoichiometric experiments which show very similar charges at oil–water and air–water interfaces across several pH units.³⁶ We note that the sum frequency and the second harmonic generation experiments probe mainly the top layer of water molecules at the surface, while the macroscopic electrokinetic experiments probe the charge inside the shear plane, within about 2.5 nm of the interface, so the two sets of experiments do not necessarily contradict each other. An explanation for these observations has recently been published that depends on the hydroxide's constraint of neighboring water molecules. An accurate description of this effect is needed to explore further the hydroxide's surface affinity.³⁷

Whether the water surface is acidic or basic is important in atmospheric chemical processes, such as absorption of carbon dioxide at the ocean surface.³⁸ Therefore this issue has prompted many computational investigations of the surface affinity of hydronium and hydroxide ions. As in the case of solvation structure, hydronium has attracted more attention, and its surface affinity has been studied using classical MD,^{39–41} empirical models,^{42,43} and AIMD.^{44,46} There is a broad agreement among these studies that hydronium weakly binds to the interfacial region behaving like an amphiphilic molecule. The calculated free energy minimum is about 1–2 kcal/mol in all three approaches, e.g., 0.7 in polarizable MD,⁴¹ 1.8 kcal/mol in the empirical valence bond model,⁴³ and 1.3 in AIMD.⁴⁶ The situation for hydroxide is less clear. Classical MD results find a slight repulsion of hydroxide from the interface,⁴⁰ the empirical model predicts a completely flat free energy profile for hydroxide,⁴⁷ while AIMD simulations indicate that it is attracted to the surface.^{44,45} The diversity of results is perhaps not surprising when one considers that even within the AIMD

simulations the solvation structure of hydroxide is quite sensitive to the chosen density functional and that several approximations are often made in practice, such as the neglect of nuclear quantum effects or the absence of counterions. Furthermore, the conditions of an AIMD simulation are very different from those of a typical experiment. For instance, the concentration of hydroxide is very high, and both the sampling and size of the system are rather limited. Nevertheless AIMD has been shown to provide an accurate description of the solvation shell of ions, and in that regard, it can be used to help improve modeling of hydroxide and resolve the differences among various empirical approaches.

Here we perform AIMD simulations of aqueous hydroxide to study the structure and polarization of the solvation waters. AIMD simulations have been previously used to study the polarization of solvation waters for various ionic systems^{48–55} but not for hydroxide. Because of its unique properties, aqueous hydroxide provides a thorough benchmark for classical models. By the same token, the polarization properties of solvation waters obtained from the AIMD simulations could provide useful insights in the development of polarizable water models. The paper is organized as follows: First, we compare the AIMD simulation results of hydroxide in water with those of classical simulations. Second, we analyze the different coordination geometries that coexist in the first hydration shell in AIMD simulations and report the dipole of waters in the first hydration shell. Finally, we consider the dipole fluctuations around the hydroxide and show that they are significantly reduced compared to bulk water.

Methods

Classical Simulations. We studied three classical models for the solvation of the hydroxide ion. The first two are based on TIP3P⁵⁶ and SPC/E⁵⁷ water models, which are rigid and nonpolarizable. They use the same hydroxide model, with a bond length of 1 Å and a charge on the hydrogen of +0.41e. We examined also hydroxide models with larger and smaller dipole moments, but these led to no essential difference in our conclusions. The third model includes electronic polarizability for all species and was recently described in ref 58. It is based on the POL3 water model. All the classical simulations were done with the NVT ensemble at the experimental density, and the temperature was controlled using the Langevin method or a Nose–Hoover chain thermostat.^{59–61} A counterion was included to neutralize the hydroxide system.

TIP3P and SPC/E Simulations. Simulations of the non-polarizable models were done using the program NAMD.⁶² Simulations of 512 water molecules at the experimental density were equilibrated initially for 10 ps at 298 K. The end point of this run was used to generate 25 starting configurations with the same positions but different velocities. These 25 configurations were equilibrated for 100 ps and then run for a further 100 ps to collect data on the structure and dielectric fluctuations. Note that this equilibration time is roughly 10 times the dielectric relaxation time

of about 10 ps. The same procedure was followed for models with 511 water molecules and a single hydroxide in the same volume, except that data were collected for 500 ps after equilibration for 100 ps. In all these simulations, the time step was 1 fs, all bonds were held rigid, and the cutoff for nonbonded interactions was 9 Å. The long-range electrostatic interactions were handled using the particle mesh Ewald (PME) method⁶³ based on a grid of 24^3 points.

POL3 Simulations. Simulations of the polarizable model were done with the toyMD program.⁶⁴ The hydroxide model is described in ref 58. It is based on an optimized fixed point charge model for hydroxide, parametrized with MP2 calculations, and used with the polarizable water model (POL3).⁶⁵ A simulation of 512 water molecules at the experimental density was split into 25 runs with different initial velocities, each was equilibrated for 20 ps, and the radial distribution functions (RDFs) and fluctuation data were then collected from each system over 40 ps. We verified with the nonpolarizable models that this much shorter simulation time is adequate for pure water, in part because we can calculate the fluctuation distributions by averaging over each molecule in the simulation. We studied also a single hydroxide with 510 POL3 waters in the same box. This system was equilibrated for 20 ps and then split into 25 runs, starting from the same positions but different velocities. Each of these 25 was equilibrated for 20 ps, and then data were collected during a further 40 ps, to give a total of 1 ns. Again, we confirmed by examining the rigid models that these equilibration times are ample. All polarizable runs used a time step of 1.25 fs, constrained the bonds to be of fixed length with the SHAKE algorithm,⁶⁶ used a cutoff for nonbonded interactions of 10.6 Å, and used the P³M method⁶⁷ on a grid of 18^3 points. Induced dipoles were calculated by a simple iterative procedure at each time step.

Ab initio Simulations. The initial coordinates for the quantum mechanics (QM) simulations were taken from equilibrated structures obtained by performing classical MD simulations in the NPT ensemble for 1 ns, at the experimental density. Two systems have been studied: (i) a water box of 64 H₂O molecules, and (ii) an aqueous hydroxide in a water box of 63 H₂O, which corresponds to a hydroxide concentration of ~1 M. A background charge was used to neutralize the hydroxide charge. This method has been found to provide accurate results even for divalent ions.⁶⁸ The calculations were performed using the Car–Parrinello MD simulations⁶⁹ with the CPMD code.⁷⁰ The CPMD equations are integrated using a velocity Verlet algorithm with a time step of 0.19 fs and using a fictitious mass of 400 au for the electrons. The valence–core interaction was described by norm-conserving Troullier–Martins pseudopotentials.⁷¹ A plane wave cutoff of 80 Ry was used. The electronic problem was solved using DFT with the BLYP exchange–correlation functional.^{22,23} Recently, this methodology was shown to describe accurately the total radial distribution function obtained in X-ray experiments.³⁰ The simulations were performed in the canonical or NVT ensemble at 298 K using a Nose–Hoover chain thermostat⁶¹ with a frequency of 1200 cm⁻¹. Two 30 ps simulations were performed for each system (water and aqueous hydroxide). The first simulation was run at a density

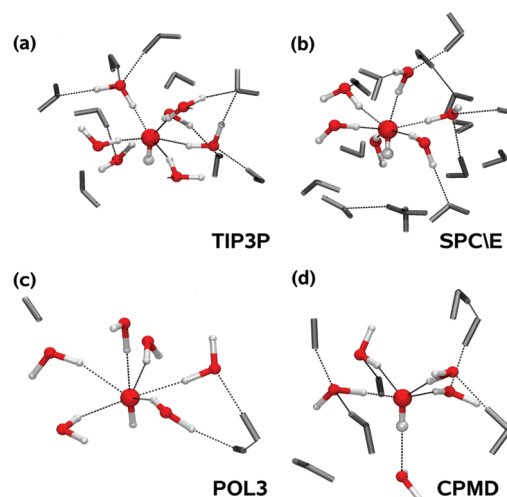


Figure 1. Coordination geometry around the hydroxide ion with different water models: (a) TIP3P, (b) SPC/E, (c) POL3, and (d) the reference AIMD calculations. Waters within 3.3 Å from the hydroxide oxygen O* are shown in CPK representation. An H-bond is counted if the distance between two heavy atoms (O* and O in this case) is less than 3.0 Å and the angle O*H*O_w is larger than 120°.

of 1.00, and the second simulation was run at a density of 1.04. The size of the periodic cubic box used was $(12.44 \times 12.44 \times 12.44 \text{ Å}^3)$ and $(12.23 \times 12.23 \times 12.23 \text{ Å}^3)$, respectively. Merz–Kollman charges⁷⁴ were computed for 200 sets of coordinates extracted from the AIMD trajectories (60 ps) and used to evaluate the magnitude of charge transfer from the hydroxide to its coordination shell. The charges were found to be insensitive to the box size. Wannier function centers^{72,73} (WFCs) were computed for each set of coordinates and used to calculate the water dipole moments.

Results

Comparison between Classical and AIMD Simulations. In Figure 1, we show representative snapshots from the classical and AIMD simulations. In the classical models the hydroxide ion is seen to be coordinated by six HB donors, while there are only four HB donors in AIMD. This qualitative picture of overcoordination in classical models is quantified in Figure 2, where we compare the RDF's and the coordination numbers obtained from the four models depicted in Figure 1 (the hydroxide atoms are denoted with a star, and water atoms are denoted with a subscript w). The three panels show the RDF's and the coordination numbers for O*–O_w (top) O*–H_w (middle), and H*–O_w (bottom). The positions of the peaks in the RDF's are different in the classical MD and AIMD simulations. In particular, shorter O*–O_w distances are observed with nonpolarizable waters (~2.6 Å), compared to (~2.8 Å) for AIMD simulations. The coordination numbers are also quite different in the two cases. As shown in the middle panel, there are approximately six HB donors in the classical MD simulations with nonpolarizable force fields but only approximately four HB donors in AIMD. Inclusion of polarization improves the classical results, but there are still sizable discrepancies especially in the H_w coordination (middle panel). Coordination of the

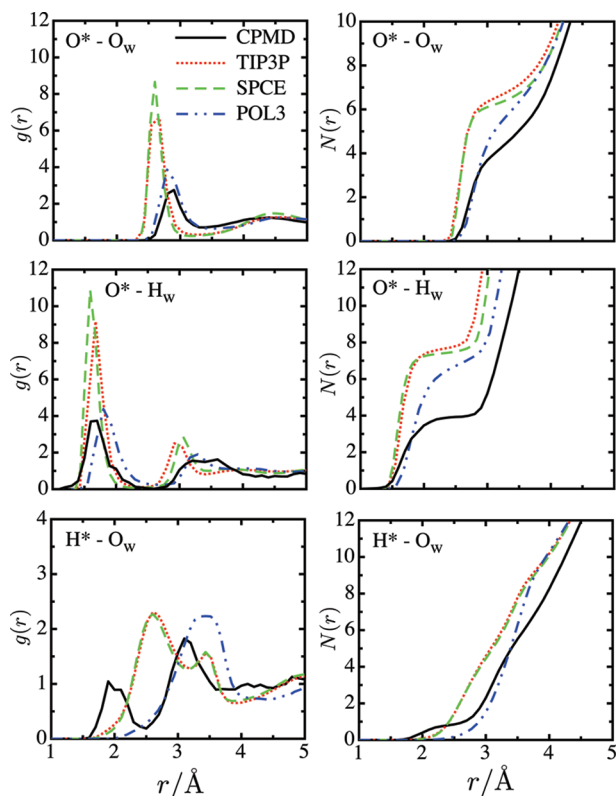


Figure 2. $O^* - O_w$ and $O^* - H_w$ partial RDF's (left) and running coordination numbers obtained by integration of the partial RDF's (right).

hydroxide hydrogen (H^*) is also different in the classical and ab initio models (bottom panel). H^* is coordinated by a water molecule during most of the simulation time in AIMD. Using a distance cutoff of 3.3 \AA, corresponding to the minimum in the $O^* - O_w$ RDF, this bond is present >60% of the simulation time. The HB made by H^* is longer compared to other HB's in the first shell, and the water exchange happens more rapidly, which points to a relatively weaker HB. In contrast, in the classical MD simulations there are no HB acceptors in the first solvation shell, and the hydroxide is coordinated only through HB donors. The differences between the classical and ab initio models with regard to the coordination numbers and the nature of HB's (i.e., donor or acceptor) indicate that the present classical models of hydroxide are not very realistic and need to be improved before use in dynamical calculations.

Coordination Geometry and Proton Transfer. The coordination geometry of aqueous hydroxide has been described previously.¹⁶ Three dominant complexes are observed in AIMD: $OH^-(H_2O)_3$, $OH^-(H_2O)_4$, and $OH^-(H_2O)_5$. The distribution among these complexes is dependent upon the concentration and the counterion used in the simulation system.^{16,19,20} We introduce the notation $(X + Y)$ to describe the complexes, where X refers to HB donors that interact with the hydroxide O^* , and Y refers to HB acceptors that interact with the hydroxide H^* . In total, six common complexes can be identified: $(3 + 1)$, $(4 + 1)$, $(5 + 1)$, $(3 + 0)$, $(4 + 0)$, and $(5 + 0)$. The probability of occurrence of these complexes is shown in Figure 3 (top). The dominance of the $(4 + Y)$ complex is clearly seen in this graph.

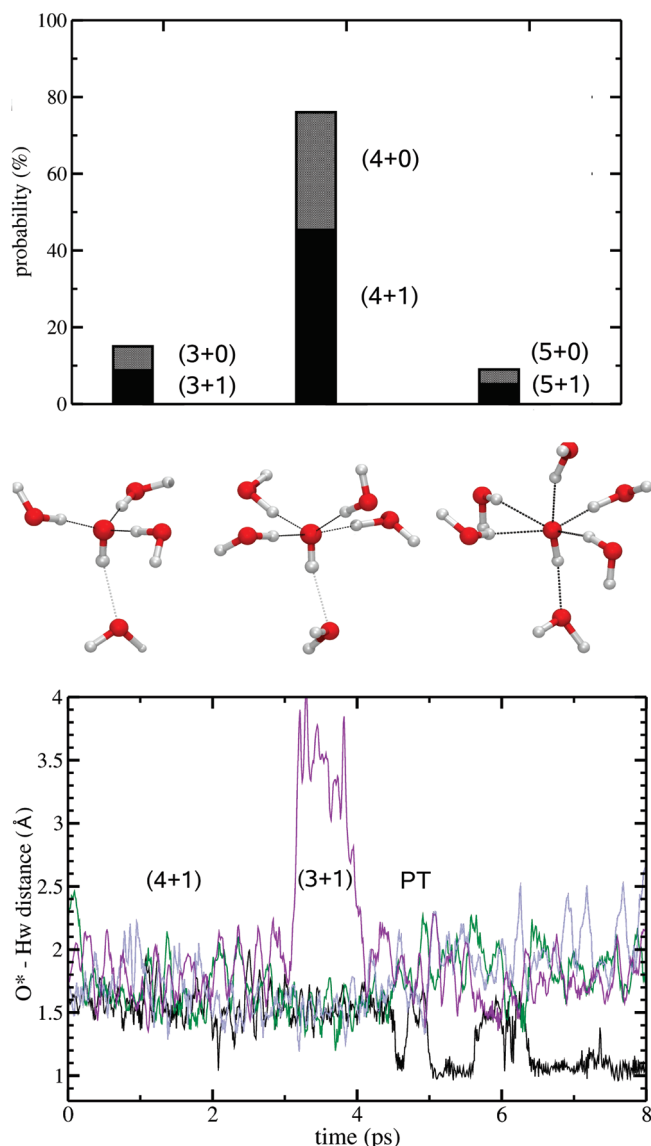


Figure 3. Histogram of the different coordination numbers in the first hydration shell obtained in the AIMD simulations (top). The probability that a water coordinates the hydroxide through the H^* is shown in black. The middle panel shows typical snapshots for the most common complexes, $(3 + 1)$, $(4 + 1)$ and $(5 + 1)$. The distance between the water ligands H_w and the hydroxide O^* as a function of time in one of the simulations (bottom).

In Figure 3 (bottom), we also show the time series for the $O^* - H_w$ distance in one of the AIMD simulations. At the start of this simulation, the hydroxide forms a stable $(4 + 1)$ complex. After ~ 3 ps, one of the HB donating waters leaves the first coordination shell, and the hydroxide forms a $(3 + 1)$ complex. From the $O^* - H_w$ distance, this water appears to come back to the solvation shell at ~ 4 ps but, in fact, it does not form an HB with O^* , hence the hydroxide remains in the $(3 + 1)$ complex. In the $(3 + 1)$ complex, the $O^* - H_w$ bonds become shorter, i.e., 1.53 ± 0.12 versus 1.70 ± 0.19 \AA in the $(4 + 1)$ complex. The smaller amplitudes in the $O^* - H_w$ vibrations indicate that the waters are also more tightly bound in the $(3 + 1)$ complex. After about 4 ps, a proton transfer (PT) event is observed to occur in this simulation. In all the AIMD simulations performed, a PT

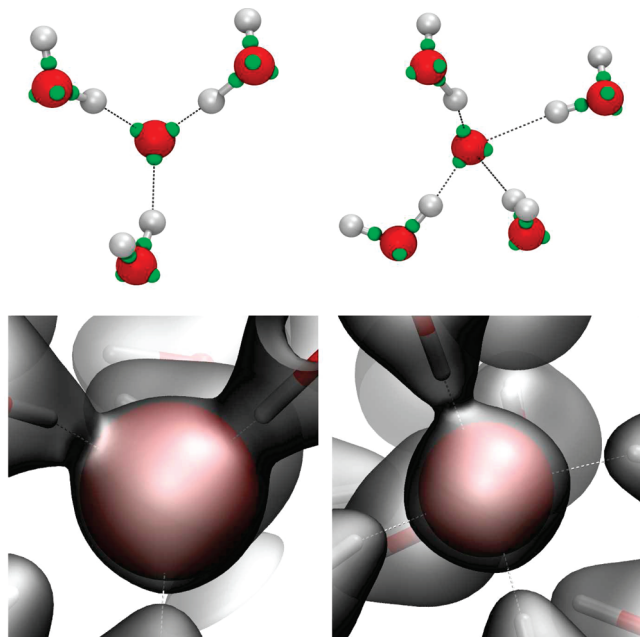


Figure 4. WFCs (top) and density isosurface of 0.05e (bottom) for the two most populated complexes in the simulations: (3 + 1) on the left and (4 + 1) on the right. The hydroxide is at the center of the images.

event has occurred within <5 ps of simulation in each case. The (3 + 1) complex has been identified as the main precursor to PT events in previous AIMD studies.¹⁶ This is also observed in our simulations, which is consistent with the presolvation concept.^{16,26} According to this concept, PT is more likely to occur in the (3 + 1) complex because the hydroxide will be transformed into a water that is already coordinated by four ligands in a tetrahedral geometry. The presence of frequent PT events has been used to rationalize the large diffusion coefficient of hydroxide relative to water.^{26,27}

An important new result of the present simulations is the finding that the O*–H_w bonds are ~0.2 Å shorter in the (3 + 1) complex compared to that of the (4 + 1) and that the amplitudes in the bond vibrations are reduced. In the next subsection, we also show that the electronic structure around the hydroxide O* is different in the (3 + 1) and (4 + 1) complexes, suggesting a new interpretation for the high reactivity of the (3 + 1) complex.

Electronic Structure. The electronic structure around the hydroxide ion has been previously investigated by Tucker et al.¹⁶ Calculations of the electrostatic potential indicated a spherical distribution of the electron density around the hydroxide O*. Using a finer evaluation of the electron density around O*, here we find that the electron density deviates slightly from a perfect spherical distribution (Figure 4). In addition, the three Wannier function centers (WFCs), corresponding to the six valence electrons present in the electron cloud around O*, point directly toward the water ligands in the (3 + 1) complex but not in the (4 + 1) complex. In the (4 + 1) complex, the most common situation occurs when two WFCs point directly toward two water molecules, and the remaining one WFC is shared between two water molecules. Inspection of the trajectories reveals

Table 1. Comparison of the Dipole Moment of Waters in the First Hydration Shell of OH[−] and Cl[−]^a

config.	dipole moment (D)	av. dist. (Å)
OH [−] <u>3</u> + 1	3.52 (0.26)	2.52
OH [−] <u>4</u> + 1	3.27 (0.42)	2.67
H ₂ O (bulk)	2.96 (0.30)	2.79
OH [−] 4 + <u>1</u>	2.86 (0.23)	3.06
Cl [−]	2.87 (0.27)	3.10

^a The solvation configuration of the hydroxide complex is listed next to the ion. The nature of water used in the calculation (i.e., HB donor or acceptor) is indicated by boldface and underlining. The bulk water and Cl[−] results are taken from refs 52 and 55. The last column gives the average distance of the hydration oxygen from the central ion or water.

that the waters directly interacting with a WFC are more likely to give a proton to the hydroxide, which indicates that they are forming the “most active” hydrogen bonds.^{7,16}

To characterize the electronic structure further, we have calculated the average dipole moment of water molecules in the first solvation shell of hydroxide using the maximally localized Wannier functions (Table 1). Here we distinguish between different solvation complexes and whether the water is HB donor or acceptor. A general observation is that the dipole moment of the HB accepting water is lower than the bulk water value, while those of HB donors are much higher than the bulk water value. The dipole moment of the acceptor is seen to be very similar to those in the hydration shell of the Cl[−] ion. As shown in the Table 1, there is a close correlation between the dipole moments of water molecules in the first solvation shell and their distance from the central ion or water. To some extent, this behavior can be traced to classical electrostatics where the polarization of water is proportional to the electric field and hence decreases with increasing distance from the ion. Charge transfer from the ion to the neighboring waters can also result in enhancement of the water dipoles, which is discussed further below.

Perhaps the most interesting result from the present AIMD simulations is that the average dipole moment of HB donors is 3.27 D in the (4 + 1) complex, which increases to 3.52 D in the (3 + 1) complex. To the best of our knowledge the water dipole moment in the (3 + 1) complex is the highest dipole moment reported for water in the hydration shell of monovalent or divalent ions. Furthermore, an analysis of the electrostatic potential indicates that the high water dipoles in the (3 + 1) complex cannot be simply explained by the lower dipole–dipole repulsion between first shell ligands, as in the case of other ions.⁵³ Instead, the observed differences in water polarization suggest that the (3 + 1) complex leads to slightly more covalent bonds that are stronger compared to the O*–H_w bonds in the (4 + 1) complex. The stronger bonds and the observed differences in the electronic structures between the two complexes are consistent with a PT mechanism that mainly involves the (3 + 1) complex, as observed in the AIMD simulations with the BLYP functional.

We have calculated the magnitude of the charge-transfer effects by computing the Merz–Kollman charges from the AIMD simulations.⁷⁴ For hydroxide in the gas phase, the charge on the O atom is −1.22e and on the H atom is +0.22e. In the aqueous phase, on the other hand, the charges

on the O and H atoms become, respectively, $-1.16e$ and $+0.32e$ for the $(4 + X)$ state and $-1.01e$ and $+0.29e$ for the $(3 + X)$ state. Thus the total charge on the hydroxide is increased from $-e$ in the gas phase to $-0.84e$ for the $(4 + X)$ state and $-0.72e$ for the $(3 + X)$ state. The corresponding charge transfers to the neighboring waters are $-0.16e$ and $-0.28e$, respectively, which are quite substantial and partly explain the increased dipole moments of waters in the hydration shell of hydroxide. Therefore, in addition to the proton transfer and induced polarization effects, the charge transfer provides another interesting phenomenon that is not described by classical models.

Fluctuations of Water Dipoles around the Hydroxide Ion. Whether the hydroxide ion is attracted to the air–water interface is a topical question. Suppression of the fluctuations of water dipoles around hydroxide has been proposed as a mechanism for its surface affinity.³⁷ Experimentally, sodium hydroxide is known to have one of the highest dielectric decrements for monovalent salts of monatomic ions. Buchner et al.⁷⁸ report that the dielectric decrement for sodium hydroxide is $20.9 \pm 0.8 \text{ M}^{-1}$ and that for sodium chloride is $15.2 \pm 0.3 \text{ M}^{-1}$. Although the suppression of fluctuations by the hydroxide is consistent with what we know from its unusual solvation structure, the quantification of the high dielectric decrement related to this structure from computer simulations is challenging. In principle, this question can be answered most directly by calculating the ensemble averages for the total dipole moment and its fluctuations in a sphere centered around a hydroxide ion. However, classical studies aimed at computing this property suggest that simulations in the order of 1–2 ns are required to reach convergence.^{75,76} Our attempt to estimate the dielectric decrement from an AIMD simulation lasting ~ 60 ps did not succeed because the present sampling time was not sufficient to obtain a statistically meaningful average that could be compared with experiments. Similarly, we note that a recent study⁷⁷ of statistical uncertainties in determination of the diffusion coefficient from AIMD simulations has concluded that ~ 500 ps would be required in ambient conditions, which is an order of magnitude longer than that used here.

In order to obtain a semiquantitative description of the reduction of dipole fluctuations, we have computed the relaxation time, τ , for the orientational autocorrelation function, $C(t)$, of O–H vectors in the simulation box. The computed orientational relaxation for waters around the hydroxide ion and for bulk water are shown in Figure 5. In the case of hydroxide, the relaxation time is found to be an order of magnitude larger than that of bulk water, which shows that the presence of hydroxide has a substantial stabilizing effect on the neighboring waters. A more intuitive picture emerges from the inspection of the trajectories in the two simulation systems. A hydroxide ion is observed to rotate slower compared to a central water molecule in the bulk system. Furthermore, the hydration waters around the hydroxide ion undergo mainly translational motion and exhibit little rotation. In contrast, the orientations of the waters around the central water change considerably during the same period. This provides a qualitative illustration of how the presence of a hydroxide ion has a stabilizing effect

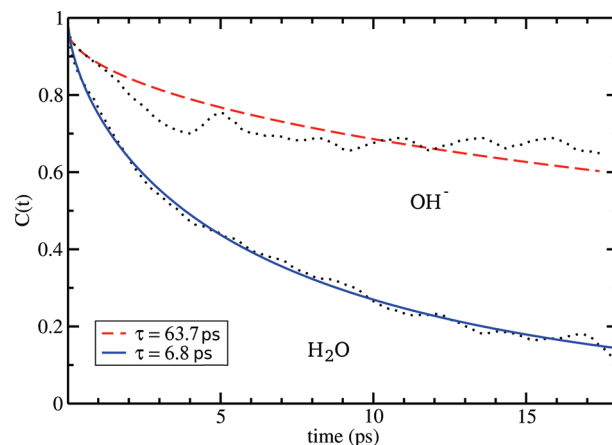


Figure 5. Comparison of the orientational correlation time (τ) of waters around a hydroxide ion (dashed line) with that of a central water in bulk (solid line). The lines are obtained by fitting, $C(t) = e^{-t/\tau}$, to the AIMD results (indicated with dots), which are calculated from 30 ps AIMD simulations. The correlation times for the two cases are given in the box.

on the neighboring water molecules and thus suppresses the dipole fluctuations in the hydroxide system relative to bulk water. These results are consistent with experimental dielectric decrements for hydroxide.⁷⁸ We are currently exploring the consequences of this effect in more detail.

Conclusions

The comparison between the classical and ab initio molecular dynamics (AIMD) simulations of aqueous hydroxide in this work indicates that the structural and electronic properties of aqueous hydroxide are not well described by current point charge models. In this regard, the newly proposed charged ring model⁷⁹ offers a considerable improvement as it predicts the four hydrogen-bond (HB) donor waters as the dominant configuration. However, one shortcoming of this model is that it fails to reproduce the weakly bound HB acceptor. Further improvement is possible, for example, by constructing a classical model that can account for the proton transfer events observed in the AIMD simulations, see, e.g., refs 80 and 81.

We have observed several proton transfer events during the AIMD simulations. In each case, proton transfer occurred not in the dominant $(4 + 1)$ complex but in the rarer $(3 + 1)$ complex. This is consistent with the presolvation concept, that is, proton transfer is more likely to occur when the hydroxide ion turns into a water molecule with a tetrahedral coordination. Calculation of the dipole moments of hydration waters has shed further light on the unique structure of the hydroxide complex. Already in the $(4 + 1)$ complex, the average dipole moment of hydration waters is substantially larger than that of bulk water, which is opposite to what happens for monovalent ions. In the case of the $(3 + 1)$ complex, the dipole moment attains a much larger value, exceeding even those in the hydration shell of a divalent ion. These unique properties make the aqueous hydroxide an ideal system for testing the polarizable force fields currently under construction.

We have also discussed how fluctuations of water dipoles around the hydroxide ion are reduced when compared to bulk

water due to the extra stability provided by the hydroxide. This effect may help to explain the affinity of hydroxide ions near the water surface and needs to be studied further with improved sampling.

Acknowledgment. This work was supported by grants from the Australian Research Council. Calculations were performed using the SGI Altix clusters at the National Computational Infrastructure (Canberra), the Australian Center for Advanced Computing and Communications (Sydney), and the Victorian Partnership for Advanced Computing (Melbourne).

References

- Atkins, P.; de Paula, J. *Atkins' Physical Chemistry*; 7th ed.; Oxford University Press: Oxford, U.K., 2002; pp 766.
- Tuckerman, M. E.; Laasonen, K.; Sprik, M.; Parrinello, M. *J. Chem. Phys.* **1995**, *103*, 150–161.
- Tuckerman, M. E.; Marx, D.; Klein, M. L.; Parrinello, M. *Science* **1997**, *275*, 817–820.
- Marx, D.; Tuckerman, M. E.; Hutter, J.; Parrinello, M. *Nature* **1999**, *397*, 601–604.
- Marx, D. *ChemPhysChem* **2006**, *7*, 1848–1870.
- Asthagiri, D.; Pratt, L. R.; Kress, J. D. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6704–6708.
- Marx, D.; Chandra, A.; Tuckerman, M. E. *Chem. Rev.* **2010**, *110*, 2174–2216.
- Agmon, N. *Chem. Phys. Lett.* **1995**, *244*, 456–462.
- Lobaugh, J.; Voth, G. A. *J. Chem. Phys.* **1996**, *104*, 2056–2069.
- Schmitt, U. W.; Voth, G. A. *J. Chem. Phys.* **1999**, *111*, 9361–9381.
- Voth, G. A. *Acc. Chem. Res.* **2006**, *39*, 143–150.
- Ando, K.; Hynes, J. T. *J. Phys. Chem. B* **1997**, *101*, 10464–10478.
- Vuilleumier, R.; Borgis, D. *J. Chem. Phys.* **1999**, *111*, 4251–4266.
- Kornyshev, A. A.; Kuznetsov, A. M.; Spohr, E. J.; Ulstrup, J. *J. Phys. Chem. B* **2003**, *107*, 3351–3366.
- Brancato, G.; Tuckerman, M. E. *J. Chem. Phys.* **2005**, *122*, 224507.
- Tuckerman, M. E.; Marx, D.; Parrinello, M. *Nature* **2002**, *417*, 925–929.
- Huckel, E. *Z. Elektrochem.* **1928**, *34*, 546–562.
- Agmon, N. *Chem. Phys. Lett.* **2000**, *319*, 247–252.
- Zhu, Z.; Tuckerman, M. E. *J. Chem. Phys. B* **2002**, *106*, 8009–8018.
- Chen, B.; Ivanov, I.; Park, J. M.; Parrinello, M.; Klein, M. L. *J. Phys. Chem. B* **2002**, *106*, 12006–12016.
- Perdew, J. P.; Wang, Y. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1992**, *45*, 13244–13249.
- Becke, A. D. *Phys. Rev. A: At., Mol., Opt. Phys.* **1988**, *38*, 3098–3100.
- Lee, C. T.; Yang, W. T.; Parr, R. G. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1988**, *37*, 785–789.
- Asthagiri, D.; Pratt, L. R.; Kress, J. D.; Gomez, M. A. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 7229–7233.
- Hamprecht, F. A.; Cohen, A. J.; Tozer, D. J.; Handy, N. C. *J. Chem. Phys.* **1998**, *109*, 6264–6271.
- Tuckerman, M. E.; Chandra, A.; Marx, D. *Acc. Chem. Res.* **2006**, *39*, 151–158.
- Chandra, A.; Tuckerman, M. E.; Marx, D. *Phys. Rev. Lett.* **2007**, *99*, 145901.
- Imberti, S.; Botti, A.; Bruni, F.; Cappa, G.; Ricci, M. A.; Soper, A. K. *J. Chem. Phys.* **2005**, *122*, 194509.
- McLain, S. E.; Imberti, S.; Soper, A. K.; Botti, A.; Bruni, F.; Ricci, M. A. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2006**, *74*, 094201.
- Megyes, T.; Balint, S.; Grosz, T.; Radnai, T.; Bako, I.; Sipos, P. *J. Chem. Phys.* **2008**, *128*, 044501.
- Marinova, K. G.; Alargova, R. G.; Denkov, N. D.; Velez, O. D.; Petsev, D. N.; Ivanov, I. B.; Borwankar, R. P. *Langmuir* **1996**, *12*, 2045–2051.
- Beattie, J. K.; Djerdjev, A. M. *Angew. Chem., Int. Ed.* **2004**, *43*, 3568–3571.
- Petersen, P. B.; Saykally, R. J. *J. Phys. Chem. B* **2005**, *109*, 7976–7980.
- Petersen, P. B.; Saykally, R. J. *Chem. Phys. Lett.* **2008**, *458*, 255–261.
- Tian, C. S.; Shen, Y. R. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 15148–15153.
- Creux, P.; Lachaise, J.; Graciaa, A.; Beattie, J. K.; Djerdjev, A. M. *J. Phys. Chem. B* **2009**, *113*, 14146–14150.
- Gray-Weale, A.; Beattie, J. K. *Phys. Chem. Chem. Phys.* **2009**, *11*, 10994–11005.
- Orr, J. C.; Fabry, V. J.; Aumont, O.; Bopp, L.; Doney, S. C.; Feely, R. A.; Gnanadesikan, A.; Gruber, N.; Ishida, A.; Joos, F.; et al. *Nature* **2005**, *437*, 681–686.
- Mucha, M.; Frigato, T.; Levering, L. M.; Allen, H. C.; Tobias, D. J.; Dang, L. X.; Jungwirth, P. *J. Phys. Chem. B* **2005**, *109*, 7617–7623.
- Buch, V.; Milet, A.; Vacha, R.; Jungwirth, P.; Devlin, J. P. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 7342–7347.
- Wick, C. D.; Kuo, I. F. W.; Mundy, C. J.; Dang, L. X. *J. Chem. Theory Comp.* **2007**, *3*, 2002–2010.
- Petersen, M. K.; Iyengar, S. S.; Day, T. J. F.; Voth, G. A. *J. Phys. Chem. B* **2004**, *108*, 14804–14806.
- Iuchi, S.; Chen, H.; Paesani, F.; Voth, G. A. *J. Phys. Chem. B* **2009**, *113*, 4017–4030.
- Kudin, K. N.; Car, R. *J. Am. Chem. Soc.* **2008**, *130*, 3915–3919.
- Mundy, C. J.; Kuo, I. F. W.; Tuckerman, M. E.; Lee, H. S.; Tobias, D. J. *Chem. Phys. Lett.* **2009**, *481*, 2–8.
- Lee, H. S.; Tuckerman, M. E. *J. Phys. Chem. B* **2009**, *113*, 2144–2151.
- Wick, C. D.; Dang, L. X. *J. Phys. Chem. B* **2009**, *113*, 6356–6364.
- Bako, I.; Hutter, J.; Palinkas, G. *J. Chem. Phys.* **2002**, *117*, 9838–9843.
- Krekeler, C.; Hess, B.; Delle Site, L. *J. Chem. Phys.* **2006**, *125*, 054305.
- Krekeler, C.; Delle Site, L. *J. Phys.: Condens. Matter* **2007**, *19*, 192101.

- (51) Whitfield, T. W.; Varma, S.; Harder, E.; Lamoureux, G.; Rempe, S. B.; Roux, B. *J. Chem. Theory Comput.* **2007**, *3*, 2068–2082.
- (52) Ikeda, T.; Boero, M.; Terakura, K. *J. Chem. Phys.* **2007**, *126*, 034501.
- (53) Bucher, D.; Kuyucak, S. *J. Phys. Chem. B* **2008**, *112*, 10786–10790.
- (54) Scipioni, R.; Schmidt, D. A.; Boero, M. *J. Chem. Phys.* **2009**, *130*, 024502.
- (55) Guardia, E.; Skarmoutsos, I.; Masia, M. *J. Chem. Theory Comp.* **2009**, *5*, 1449–1453.
- (56) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926–935.
- (57) Berendsen, H. J. C.; Grigera, J. R.; Straatsma, T. P. *J. Phys. Chem.* **1987**, *91*, 6269–6271.
- (58) Vacha, R.; Horinek, D.; Berkowitz, M. L.; Jungwirth, P. *Phys. Chem. Chem. Phys.* **2008**, *10*, 4975–4980.
- (59) Nose, S. *J. Chem. Phys.* **1984**, *81*, 511–519.
- (60) Hoover, W. G. *Phys. Rev. A: At., Mol., Opt. Phys.* **1985**, *31*, 1695–1697.
- (61) Martyna, G. J.; Klein, M. L.; Tuckerman, M. *J. Chem. Phys.* **1992**, *97*, 2635–2643.
- (62) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kale, L.; Schulten, K. *J. Comput. Chem.* **2005**, *26*, 1781–1802.
- (63) Darden, T.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993**, *98*, 10089–10092.
- (64) Gray-Weale, A. *Program toyMD*; Monash University: Victoria, Australia, 2010; <https://confluence-vre.its.monash.edu.au/display/toyMD/>.
- (65) Caldwell, J. W.; Kollman, P. A. *J. Phys. Chem.* **1995**, *99*, 6208–6219.
- (66) Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *23*, 327–341.
- (67) Hockney, R. W.; Eastwood, J. W. *Computer Simulation Using Particles*; Institute of Physics: Bristol, U.K., 1981.
- (68) Todorova, T.; Hunenberger, P. H.; Hutter, J. *J. Chem. Theory Comput.* **2008**, *4*, 779–789.
- (69) Car, R.; Parrinello, M. *Phys. Rev. Lett.* **1985**, *55*, 2471–2474.
- (70) Hutter, J.; Alavi, A.; Deutch, T.; Bernasconi, M.; Goedecker, S.; Marx, D.; Tuckerman, M.; Parrinello, M. *Technical Report*; MPI fur Festkorperforschung and IBM Zurich Research Laboratory: Zurich, Switzerland, 1995–1999.
- (71) Troullier, N.; Martins, J. L. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1991**, *43*, 1993–2006.
- (72) Marzari, N.; Vanderbilt, D. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1997**, *56*, 12847–12865.
- (73) Silvestrelli, P. L.; Parrinello, M. *J. Chem. Phys.* **1999**, *111*, 3572–3580.
- (74) Singh, U. C.; Kollman, P. A. *J. Comput. Chem.* **1984**, *5*, 129–145.
- (75) Ronne, C.; Thrane, L.; Astrand, P. O.; Wallqvist, A.; Mikkelsen, K. V.; Keiding, S. R. *J. Chem. Phys.* **1997**, *107*, 5319–5331.
- (76) Heinz, T. N.; van Gunsteren, W. F.; Hunenberger, P. H. *J. Chem. Phys.* **2001**, *115*, 1125–1136.
- (77) Kuo, I. F. W.; Mundy, C. J.; McGrath, M. J.; Siepmann, J. I. *J. Chem. Theory Comp.* **2006**, *2*, 1274–1281.
- (78) Buchner, R.; Hefter, G.; May, P. M.; Sipos, P. *J. Phys. Chem. B* **1999**, *103*, 11186–11190.
- (79) Ufimtsev, I. S.; Kalinichev, A. G.; Martinez, T. J.; Kirkpatrick, R. *J. Chem. Phys. Lett.* **2007**, *442*, 128–133.
- (80) Billeter, S. R.; van Gunsteren, W. F. *J. Phys. Chem. A* **1998**, *102*, 4669–4678.
- (81) Pomes, R.; Roux, B. *Biophys. J.* **2002**, *82*, 2304–2316.

JCTC

Journal of Chemical Theory and Computation

Quantum Mechanical and Quantum Mechanical/Molecular Mechanical Studies of the Iron–Dioxygen Intermediates and Proton Transfer in Superoxide Reductase

Patrick H.-L. Sit,^{*,†,‡} Agostino Migliore,^{†,‡,§} Ming-Hsun Ho,^{†,‡} and Michael L. Klein^{†,||}

Department of Chemistry, University of Pennsylvania, Philadelphia, Pennsylvania 19104-6323, School of Chemistry, Tel Aviv University, Tel Aviv 69978, Israel, and Institute for Computational Molecular Science, Temple University, Philadelphia, Pennsylvania 19130

Received November 14, 2009

Abstract: Classical and quantum-chemical computations are employed to probe the reaction intermediates and proton-transfer processes in superoxide reductase (SOR) from *Desulfoarculus baarsii*. Ab initio studies of the SOR active site, as well as classical and QM/MM MD simulations on the overall enzymatic reaction, are performed. We explore the use of a Hubbard U correction to standard density functional theory (DFT) in order to obtain a better description of the strongly correlated d electrons in the transition-metal center. The results obtained from the standard and Hubbard- U -corrected DFT approaches are compared with those obtained using different hybrid-DFT functionals. We show that the Hubbard U correction gives a significant improvement in the description of the structural, energetic, and electronic properties of SOR. We establish that adopting the Hubbard U correction in the QM/MM approach leads to increased accuracy with essentially no additional computational cost. Our results suggest that Lys⁴⁸ is one of the likely sources of the first proton donation to the superoxide, either directly or through an interstitial water molecule. Our QM/MM calculations highlight the important role of the interactions and hydrogen-bond network created by the imidazole rings of the His ligands and the internal water molecules. Whereas the hydrogen-bonding pattern due to internal waters can facilitate the protonation event, the interactions with the His ligands and the hydrogen bonds with water can stabilize the dioxygen ligand in a side-on conformation, which, in turn, prevents the immediate proton transfer from Lys⁴⁸, as indicated by recent experimental studies.

Introduction

The interaction of dioxygen species with transition-metal-containing proteins is important in many biological processes including transport, metabolism, respiration, and cell protection. Adventitious reduction of O₂ by redox-active enzymes can yield potentially harmful species such as superoxide, peroxide, hydroperoxide, and hydroxyl radicals, which are inevitably encountered by organisms exposed to molecular

oxygen.¹ In fact, it has been established that both aerobic² and anaerobic³ organisms developed specific mechanisms to protect themselves from oxygen radical species.

The toxicity of the molecular oxygen radicals present on the surface of Earth is a major topic in modern biology and biochemistry, with evident medical significance.⁴ For instance, high levels of superoxide (O₂^{•−}) species are implicated in a number of diseases and neurological disorders,⁵ including diabetes,⁶ Parkinson's^{7,8} and Alzheimer's^{9,10} diseases, and death and tissue damage following a stroke or heart attack.¹¹ It also seems that damage to DNA due to the superoxide radical induces mutations leading to some types of cancer.^{12,13}

* Corresponding author e-mail: sit@sas.upenn.edu.

† University of Pennsylvania.

‡ These authors contributed equally to this work.

§ Tel Aviv University.

|| Temple University.

In the past three decades, research in the field of oxygen toxicity has been focused mainly on three issues:⁴ (i) identification of toxic species, biological targets, and molecular mechanisms of oxidative stress; (ii) understanding of the nature and operation of the antioxidant machinery employed by living organisms for an appropriate balance between generation and scavenging of oxygen radicals in a transitory (anaerobes) or continuous (aerobes) way; and (iii) determination of the effects of oxidative stress on aging and the above-mentioned diseases, as well as the identification and design of suitable therapeutic strategies.

Until the 1990s, the scavenging of superoxide by dismutation to molecular oxygen and hydrogen peroxide, catalyzed by the enzyme superoxide dismutase (SOD),² was considered to be the only biological mechanism for superoxide detoxification. However, it has recently been discovered³ that some anaerobic bacteria and archaea, which might be accidentally exposed to molecular oxygen, protect against harmful oxygen compounds not only by means of SOD, but also through a different class of metalloenzymes, named superoxide reductases (SORs).³ Such enzymes selectively reduce O_2^- species to hydrogen peroxide, without formation of molecular oxygen as a byproduct.⁵ SORs show great efficiency in scavenging O_2^- . Furthermore, by shuttling electrons between auto-oxidizable soluble redox proteins and the superoxide, SORs can simultaneously eliminate O_2^- and the source of its production, also shutting off transient superoxide production without the need for sophisticated regulatory systems.⁴ Such features can play a crucial role in the future development and utilization of SOR mimics,¹⁴ as effective alternatives to SOD mimics, in the therapeutic treatment of the above-mentioned diseases.

The above issues have recently attracted increasing attention within the fields of biochemistry, biology, and medicine. It has been generally suggested that two protonation steps are essential to convert superoxide to hydrogen peroxide. However, the source of the protons, as well as the relation between protonations and the reduction process, are still under debate. Understanding of the SOR mechanism, and thus its control, requires a deep theoretical analysis of the necessary protonation steps and related electron-transfer processes, with special attention to the pertinent energetics and its connection to subtle structural rearrangements. In this work, we present a computational study of the SOR enzymatic mechanism that aims at maximizing the predictive capabilities through the combined use of different theoretical—computational techniques.

Specifically, we first perform a detailed quantum-mechanical (QM) study of the enzyme active site. Next, the behavior of the overall protein system is investigated by means of molecular dynamics (MD) simulations, which are able to capture the most probable scenarios for the two protonation steps involved in the SOR mechanism. Finally, a more accurate description of the overall system is achieved through the quantum-mechanical/molecular-mechanical (QM/MM)^{15,16} approach, where the accuracy of the QM treatment of the enzyme active site can be combined with the speed of the MM computation of the protein framework.

The geometry and energetics of the SOR active site are studied using various hybrid-density functional theory (DFT) schemes and a DFT approach where the exchange-correlation (XC) functional is improved through the introduction of a self-consistent^{17,18} Hubbard- U ¹⁹-corrected electron–electron interaction term. In fact, although DFT computational approaches generally offer the best compromise between accuracy and feasibility for investigating highly correlated many-electron systems,²⁰ they can still suffer from shortcomings related to the approximate character of any currently available XC functional. The use of disparate DFT implementations permits some conclusions to be reached about geometry and electronic structure features of the active site that are robust against the individual computational schemes. This holds at each stage of the SOR enzymatic mechanism, which is characterized by two protonation steps and different scenarios for the pertinent protonation paths. From a computational point of view, our results validate the adopted Hubbard- U -corrected DFT approach and support its use in QM/MM investigations. From a chemical point of view, our results help clarify the nature of the protonation steps, the nonproduction of iron–oxo radicals, and the role of the interstitial water that forms hydrogen bonds between the crucial residue Lys⁴⁸ (see below) and the dioxygen species.

Aims and Methods

SOR Structure and Enzymatic Mechanism: Open Questions. The SOR structure, including the dioxygen moiety, was drawn from the PDB file with the code 2J13.¹ It corresponds to the mutant E114A (Glu¹¹⁴ → Ala¹¹⁴) of SOR from the sulfate-reducing bacterium *Desulfoarculus baarsii*.²¹ The mutation does not alter the overall protein structure, as reported in ref 21 and confirmed by our simulations, but stabilizes the iron–dioxygen complexes relative to the wild-type enzyme. It has been suggested²¹ that the mutation stabilizes the reaction intermediates by avoiding assisted release of hydrogen peroxide by Glu¹¹⁴. The structure of the native reduced enzyme was drawn from the PDB file with the code 2J12.²¹

The asymmetric unit in SOR-E114A crystals includes four monomers, denoted A–D in Figure 1A. Diffraction data collected upon soaking with H_2O_2 are consistent with the formation of end-on (η^1) peroxide species in monomers B–D, whereas monomer A did not show reactivity.²¹ The SOR catalytic domain displays an immunoglobulin-like fold.^{22,23} The active site is made up of an iron atom coordinated to four equatorial histidines and one axial cysteine (Figure 1B), with a vacant sixth coordination site that is exposed to the solvent and available for superoxide binding.²⁴

The SOR catalytic cycle proposed in ref 21 is illustrated in Figure 2. The overall reaction starts with the superoxide binding to a reduced penta-coordinated active site (1) to form the iron–dioxygen complex (2). The first protonation leads to the formation of the iron–hydroperoxo complex (3). After the second protonation of the proximal oxygen atom and electron transfer from the iron to the dioxygen, hydrogen peroxide is formed, and the active site is in its oxidized state

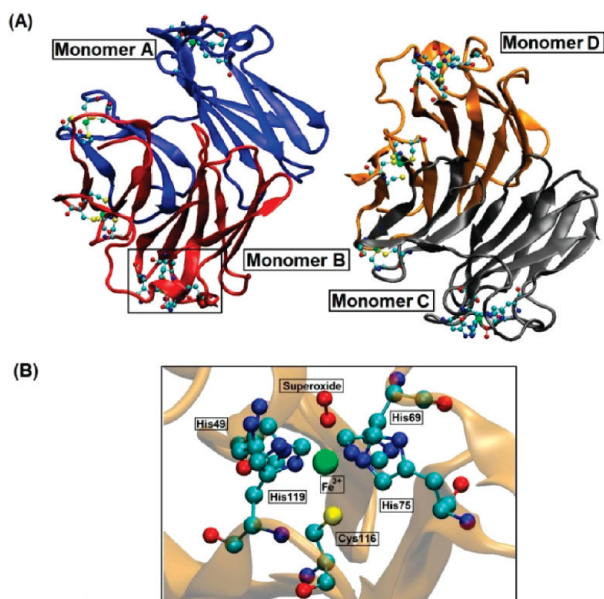


Figure 1. (A) SOR enzyme tetramer. (B) Inset illustrating the active site of monomer B with bound superoxide.

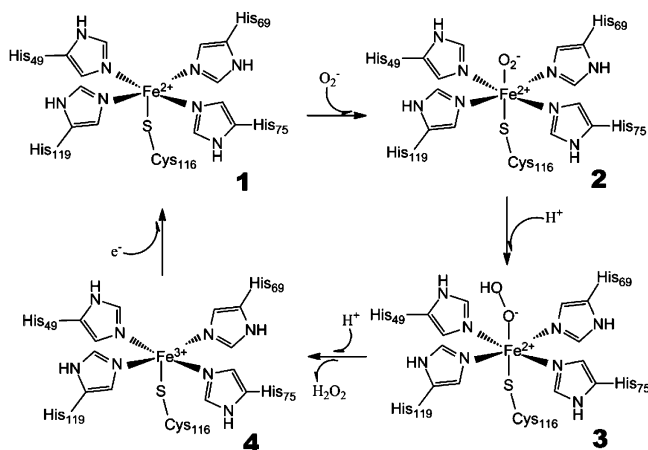


Figure 2. Proposed mechanism for superoxide reduction by SOR.

(4). Finally, the active site is regenerated to its reduced state (1) by an external reducing agent.

The unprotonated peroxy species at step 2 (Figure 2) could correspond to a high-spin side-on (η^2) configuration, as suggested by experiments,^{25–27} but this has not been confirmed by theoretical studies.^{21,28} In iron–dioxygen complexes, the total spin is given by the sum of the electron spins on iron and dioxygen. Evidence indicates that the total spin state is a decisive factor in modulating the strength of the Fe–O and O–O bonds in iron–dioxygen complexes.^{5,29} In fact, some studies have suggested that SORs involve a high-spin ($S = 5/2$) state of the iron,^{25–27} contrary to many heme-containing catalysts that involve low-spin iron states and promote the cleavage of the O–O bond. However, other studies have shown that the ground state corresponds to an intermediate ($S = 3/2$)³⁰ or low ($S = 1/2$)²⁸ spin state. Indeed, the relative stability of the active-site spin states is expected to be strongly affected by the protonation state and the environment. Furthermore, finely tuned electron donation by the Cys¹¹⁶ ligand is expected to precisely adjust the strength

of the Fe–O bond.⁵ Thus, an accurate theoretical study of the electronic distribution in the active site and the corresponding spin state, for the initially unprotonated and subsequent differently protonated intermediates of the iron–dioxygen complex, is crucial to the understanding of the dynamics of the catalytic cycle. In this way also, the connection between the protonation state of the enzyme intermediate and the spin state of the SOR–peroxide complex can be determined.

According to ref 21, the X-ray models of SOR monomers B–D show end-on iron–dioxygen species in three different conformers. Moreover, previous DFT calculations on model SOR active sites based on the X-ray structures favored high-spin end-on (η^1) hydroperoxy species protonated at the distal oxygen, which is consistent with the analysis of pulse-radiolysis data²¹ and is confirmed by the present work. In fact, our DFT calculations allow the energy ordering of the various spin states of the enzyme active site to be established consistently, in correspondence with the active site's different protonation states.

Several issues emerging from recent experimental and theoretical analyses, and partially sketched above, need to be addressed or better understood: (i) the effective proton source in each protonation step; (ii) the exact role of the water molecules and Lys⁴⁸, as well as the conformation of the flexible loop bearing the latter, in those steps; (iii) the nuclear configuration and electronic structure of the iron–dioxygen species, in relation to the protonation state and protein environment; and (iv) the role of the electron donation by the trans thiolate ligand. A deeper understanding of the above points and more predictive capability of the catalytic activity require an accurate QM study of the active site. The effects of the protein and solvent environment and the protein dynamics during the enzymatic reaction can be appreciated using classical molecular dynamics (MD) and the popular composite QM/MM approach.^{15,16} The latter combines the accuracy of the QM treatment of the active site with a classical force-field-based molecular mechanical (MM) treatment of the protein.

QM Study of SOR Active Site. The first focus is on a detailed DFT study of the SOR active site at the different stages of the enzymatic catalytic activity, corresponding to zero, one, and two additional H⁺ ions in the active zone. After the pruning of the surrounding protein, we obtained the Fe^{II}(Im₄MeS) model system (Figure 3), where Cys¹¹⁶ was replaced by a methyl thiolate and the four His ligands were replaced by four neutral imidazole rings. Only the oxygen and hydrogen atoms were allowed to move in most of our geometry optimizations of the active site.

Because of the presence of the imidazole rings and the transition metal, electron correlation plays an important role in determining the electronic properties of the system. Hence, DFT offers the best compromise between accuracy and feasibility for a comprehensive study of the enzyme active site. Furthermore, DFT allows the quality of the protein pruning to be checked through calculations on larger atomic models that are unmanageable for extensive study with post-HF approaches, which is also a crucial point in view of the QM/MM calculations discussed later in this article.

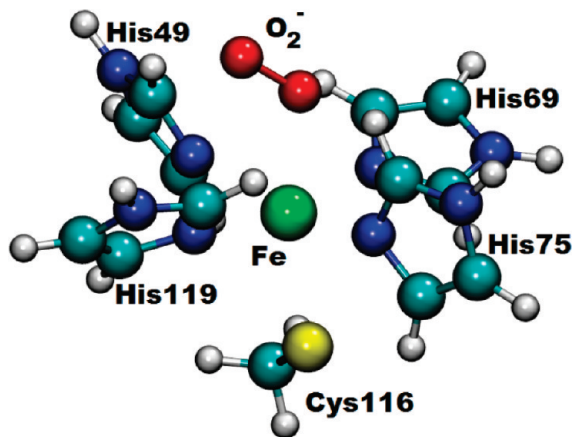


Figure 3. Model of the SOR active site, used in the DFT calculations, including the superoxide moiety (red), iron (green), sulfur (yellow), and imidazole rings (N in blue and C in cyan).

Disparate DFT schemes were adopted, including both plane-wave (PWscf code from the Quantum-Espresso distribution³¹) and atomic (NWChem package³²) basis set approaches. PWscf calculations use the PBE functional³³ in conjunction with Vanderbilt ultrasoft pseudopotentials³⁴ to represent the atomic cores. The energy cutoffs for the plane-wave expansions of the Kohn–Sham spin orbitals and the electron charge density are 30 and 240 Rydberg (Ry), respectively. The self-consistent PBE + U computational scheme^{17,18} described in the next subsection was also employed to provide an improved description of the strongly correlated electron charge around the transition-metal center.

The quantum chemical calculations using localized basis sets were performed with various hybrid XC functionals and, thus, with different amounts of exact Kohn–Sham exchange. The consistency of the results supports the robustness of our conclusions (e.g., as to the energy ordering of the intermediate- and high-spin states), against the approximations differently affecting any currently available XC functional, and especially the unphysical self-interaction arising from the fact that the approximations to the unknown exact functional are independent of the electrostatic repulsion term. Indeed, hybrid XC functionals are conceived in such a way that they cover part of the electron self-interaction error and its long-range correlation effects in a balanced manner.³⁵ As a result, they often perform better than DFT schemes specifically corrected for self-interaction. Furthermore, correcting for the residual self-interaction error does not necessarily improve the ability to reproduce various molecular properties.³⁵ Specifically, we employed the B3LYP,^{36,37} PBE0,³⁸ B97-3,³⁹ and Becke half-and-half⁴⁰ (here denoted by BHH) XC functionals, along with the 6-31+G* Pople-style basis set for the superoxide radical anion and the 6-31G* basis set for the other atoms. In fact, our computational tests indicate that the inclusion of diffuse basis functions has significant effects only regarding the O_2^- moiety.

Ab Initio PBE + U Scheme. The appropriate description of the electron–electron interactions, and thus of their correlations, is a major issue in all quantum chemical calculations. Within DFT, the above-mentioned shortcomings can significantly affect the resulting electron density. An

improved description of the valence electron density can be achieved by adding an orbital-dependent correction functional E_U , adapted from the Hubbard model,¹⁹ to the standard XC functional E_{DFT} . According to the Hubbard- U -corrected DFT approach, a few localized orbitals (specifically, the d-like orbitals around the iron center for the system under consideration) are selected, and the corresponding electron correlation is treated in a special way. The magnitude of E_U , and hence the size of the correction term, is controlled by the U parameter, which gives the strength of the screened on-site Coulomb interactions. The total energy functional is written as

$$E_{DFT+U}[n] = E_U[\{n_{mm'}^{Is}\}] + E_{DFT}[n] \quad (1)$$

where I is the atomic site that experiences the Hubbard-like interaction, s is the electronic spin, m is the magnetic quantum number, and \mathbf{n}^{Is} is the atomic orbital occupation matrix. \mathbf{n}^{Is} describes the degrees of freedom associated with the strongly correlated electrons, on which the U interaction term acts. A recently developed theoretical method²¹ offers a self-consistent recipe for evaluating the U parameter. Within a rotationally invariant formalism, the U value appropriate for a given system is derived from a linear-response approach that is internally consistent with the definition of the occupation matrix.¹⁸ The resultant expression of the corrective energy functional is¹⁸

$$E_U[\{n_{mm'}^{Is}\}] = \frac{U}{2} \sum_{I,s} \text{Tr}[\mathbf{n}^{Is}(1 - \mathbf{n}^{Is})] \quad (2)$$

The functional in eq 2 introduces a penalty, tuned by the U parameter, for partial occupations of the localized orbitals, thus favoring fully occupied or empty spin orbitals. In fact, a positive term is added to the energy in the presence of fractionally occupied orbitals. Yet, the interaction parameter U used in the self-consistent calculations can be determined essentially from first principles,⁴¹ and it need not be treated as an empirical fitting parameter. Note that the PBE + U computational scheme outlined above can allow a substantial self-interaction correction as a result of a more appropriate description of the strong electronic correlation⁴² due to the presence of the iron and the radical anion.

QM/MM Simulations. QM/MM simulations are performed here with the CP2K code.⁴³ The enzymatic system is divided into two parts: (i) the active-site region, which is treated at the QM level with the Hubbard- U -corrected PBE functional, and (ii) the protein and the surrounding water, which are treated at the MM level, using the AMBER force field. In detail, the QM part of the system includes Fe²⁺, its ligands (His⁴⁹, His⁶⁹, His⁷⁵, His¹¹⁹, and Cys¹¹⁶), Lys⁴⁸, the superoxide anion, and the water molecules within a radius of 6 Å from the latter. The unit-cell size for the QM system is 18 Å × 22 Å × 22 Å. Goedecker–Teter–Hutter pseudopotentials^{44,45} are used to describe electron and ion interactions. Unlike other DFT packages, CP2K uses dual Gaussian-type and plane-wave basis sets.⁴³ We use a basis set of double- ξ quality for Fe and triple- ξ basis sets for the other atoms. A plane-wave cutoff of 280 Ry is chosen to describe the electronic structure of the system. Link atoms

Table 1. DFT Spin-State Energies (kcal/mol) of the Native Reduced SOR^{a-c}

spin state	method					
	PBE	PBE + <i>U</i>	B3LYP	PBE0	B97-3	BHH
low (<i>S</i> = 0)	23.52	58.65	38.67	50.43	43.26	62.56
intermediate (<i>S</i> = 1)	15.68	34.81	21.40	31.47	27.72	34.81

^a Plane-wave calculations used the PBE XC functional and its self-consistent Hubbard *U* correction (*U* = 6.4 eV), denoted by PBE + *U*. ^b Hybrid-DFT evaluations used the 6-31G* basis set and the indicated hybrid functionals. ^c Energies were measured with respect to the high-spin (*S* = 2) state.

are applied, and the interaction between QM and MM regions is calculated using the procedure described by Laino et al.⁴⁶ The Born–Oppenheimer technique is used to propagate the atoms in the quantum region. The time step for the MD simulation is 0.5 fs. The system is coupled with a Nosé–Hoover thermostat^{47,48} at a 100-fs frequency to achieve constant-temperature simulations.

Results and Discussion

Spin-State Energies of the Native Reduced Enzyme. As mentioned above, the employed minimal model is Fe^{II}(Im₄MeS), and its coordinates were obtained by pruning of the protein structure from the PDB file 2JI2. Monomer B of the protein structure was chosen for this study. Note that the superoxide is absent in the native reduced form of the protein. The dangling bonds were saturated with H atoms, which were then relaxed while keeping the coordinates of the heavy atoms fixed as in the 2JI2 X-ray structure. Table 1 shows the energies of the model system in the low- and intermediate-spin states relative to the energy of the high-spin state. All of the calculations, using both plane-wave and atomic basis sets with disparate XC functionals, give the energy order high- < intermediate- < low-spin state. This conclusion is robust against the quality of the XC functional and the kind of basis set. The use of the Hubbard *U* correction to the PBE functional stabilizes the high-spin state relative to the other two spin states. Indeed, magnetic resonance experiments²⁴ indicate that the high-spin state is strongly favored. From a computational point of view, the increased stability of the high-spin state arising from the special treatment of the strongly correlated electrons with the additional Hubbard *U* electronic interaction matches the results from various hybrid XC functions. The latter generally improve the description of electronic properties, as compared to pure DFT schemes, for the following reasons: (i) The delocalization effects in the exchange are better described, thereby reducing the exaggeration of the nondynamic electron correlation. (ii) Both nondynamic and dynamic electron correlations are heeded in a more realistic and effective way, as a consequence of the fitting against experimental data.³⁵ From our data in Table 1 we conclude that the adopted Hubbard *U* correction to the PBE approximation improves the description of electron correlation effects comparably to hybrid-DFT approaches and in line with the expectations from the currently available experimental evidence. This is accomplished essentially without additional computational

Table 2. Structure and Energetics of the Model O₂⁻-Bound SOR Active Site in Its Three Possible Spin States, as Obtained from DFT Plane-Wave Calculations, at the PBE and PBE + *U* Levels, and from Hybrid-DFT Evaluations, Using the Indicated Functionals along with the 6-31+G* Basis Set for the Superoxide Radical Anion and the 6-31G* Basis Set for the Remaining Atoms

method	spin state	d _{Fe-proximal O} (Å)	d _{O-O} (Å)	∠ _{Fe-O-O} (deg)	energy (kcal/mol)
PBE	<i>S</i> = 1/2	1.83	1.38	128.9	7.49
	<i>S</i> = 3/2	1.85	1.37	129.2	-0.19
	<i>S</i> = 5/2	1.91	1.38	142.8	0.00
PBE + <i>U</i>	<i>S</i> = 1/2	2.08	1.35	127.5	45.36
	<i>S</i> = 3/2	2.30	1.35	122.5	-0.41
	<i>S</i> = 5/2	2.31	1.35	122.9	0.00
B3LYP	<i>S</i> = 1/2	1.95	1.34	128.8	21.07
	<i>S</i> = 3/2	2.20	1.34	125.5	0.33
	<i>S</i> = 5/2	2.18	1.34	128.0	0.00
PBE0	<i>S</i> = 1/2	1.96	1.32	127.7	34.37
	<i>S</i> = 3/2	2.17	1.32	125.2	0.12
	<i>S</i> = 5/2	2.17	1.32	126.7	0.00
B97-3	<i>S</i> = 1/2	1.97	1.32	128.1	30.90
	<i>S</i> = 3/2	2.23	1.33	126.9	23.48
	<i>S</i> = 5/2	2.28	1.32	126.4	0.00
BHH	<i>S</i> = 1/2	1.76	1.37	127.6	75.22 (SC = 13%) ^a
		1.73	1.37	127.7	71.02 (SC = 88%) ^a
	<i>S</i> = 3/2	2.25	1.30	125.7	37.24
	<i>S</i> = 5/2	2.23	1.30	125.9	0.00

^a SC stands for spin contamination.

cost relative to standard DFT, which supports an effective use of the PBE + *U* computational scheme in the QM/MM study of the enzyme.

Structure and Energetics of the Iron–Dioxygen Complex in Different Spin States. The minimal model, including the SOR active site and the superoxide species (see Figure 3), is derived from the X-ray structure of the SOR-superoxide intermediate in the monomer B of the mutant SOR-E114A.²¹ We carried out a geometry optimization for each of the three spin states, by relaxing the superoxide and the hydrogen atoms while keeping the other atoms fixed to their experimental positions. Some important geometric parameters obtained in this work at different computational levels are reported in Table 2. The ab initio *U* value obtained for the iron–dioxygen complex is 7.2 eV. This value is similar to the ones adopted in recent electronic structure calculations on the solvated ferrous–ferric redox couple.⁴² The spin-state relative energies are also shown. The PBE, PBE + *U*, B3LYP, and PBE0 computational levels provide very similar energies for the intermediate-spin (*S* = 3/2) and high-spin (*S* = 5/2) states (as compared with *k_BT* at room temperature), whereas the low-spin state is significantly higher in energy. This energy scheme can be attributed to the exchange interaction, which favors parallel spins for the electrons around Fe. In particular, the energies of the *S* = 3/2 and *S* = 5/2 states are close because they are expected to differ by a spin flipping of the outer electron on the superoxide, with the iron remaining in its high-spin state. This result is in contrast to recent QM and QM/MM studies of related iron–oxo systems such as cytochrome P450^{49,50} in which the iron is found to be in low-spin states. Spin-state energetics is expected to depend strongly on the active-site local environment. Moreover, native reduced SOR is experimentally shown to be in a high-spin state,²⁴ in

agreement with our calculation. Therefore, it is not surprising that the SOR active site iron retains its high-spin character when the correct local environment is considered. In fact, the exchange interaction among the electrons around Fe is much higher than that with the electron charge localized on the O_2^- moiety. On the other hand, for the same reason, the above-mentioned spin flipping will only slightly affect the energy of the iron–dioxygen complex, in favor of its high-spin state. This feature is accurately reproduced by the hybrid B3LYP and PBE0 functionals. Also, the PBE and PBE + U computational schemes are able to grasp the near-degeneracy of the $S = 3/2$ and $S = 5/2$ states. Moreover, all of these functionals catch the antiferromagnetic nature of the electron spins on the iron and on O_2^- in the intermediate-spin state, although to the detriment of an appreciable spin contamination (SC).⁵¹ In other terms, such functionals are able to yield a correct antiferromagnetic character of the charge distribution, although they lead to an overall density affected by SC because of the use of a single-determinant DFT approach and of the approximations in the XC functionals. On the contrary, against the expectations based on the exchange interactions, the intermediate-spin states obtained from the BHH and B97-3 calculations do not correspond to a high-spin Fe (see Table 4, below). As a consequence, these functionals lead to a more appreciable energy difference between the $S = 3/2$ and $S = 5/2$ states, as compared with the B3LYP and PBE0 functionals.

In any case, the main conclusions about the energy ordering of the three spin states achieved in this work appear robust against the use of different XC functionals and their inability to obtain SC-free results for the antiferromagnetic electron density corresponding to a different spatial localization of the spin-up and spin-down electrons around the Fe– O_2 complex. The robustness of the main conclusions against the SC problem is exemplified in Table 2 for the BHH results on the low-spin complex. In fact, the different amounts of SC explored through the BHH functional do not affect the energy order of the electronic spin states. It is worth noting that the BHH calculation with higher SC brings about a lower energy, in agreement with the general expectation that spin-unrestricted techniques can give qualitatively correct energies but wrong densities.⁵¹ Note also that the results regarding the high-spin state (which is used in the molecular dynamics studies) are essentially unaffected by SC.

On the other hand, further analyses are required to conclusively establish accurate values of the relative energies in Table 2. Moreover, contrasting results from the existing literature claim that the intermediate-spin state³⁰ or the high-spin one^{25–28,52} is more stable. The exact description of the antiferromagnetic nature of the electron spin densities around iron and superoxide in the $S = 3/2$ state would require accurate multireference calculations,⁵³ with a severe limitation of the accessible size of the model system, or use of the unknown exact XC functional, which would be able to cover all of the contributions to the nondynamic and dynamic electron correlations.³⁵

The low-spin state is consistently endowed with the highest energy. The difference in energy with the high-spin state is similarly appreciated by the PBE + U scheme and the hybrid

Table 3. Structure and Energetics of a Larger Model System Including the Histidine Beta Carbons^a

method	spin state	d_{Fe-S} (Å)	$d_{Fe-proximal O}$ (Å)	d_{O-O} (Å)	\angle_{Fe-O-O} (deg)	energy (kcal/mol)
PBE	$S = 3/2$	2.40	1.83	1.39	126.3	−0.35
	$S = 5/2$	2.56	1.84	1.37	155.9	0.00
PBE + U	$S = 3/2$	2.58	2.26	1.35	125.7	−0.41
	$S = 5/2$	2.56	2.31	1.35	125.9	0.00
B3LYP ^b	$S = 3/2$	2.51	2.12	1.34	127.3	−0.82
	$S = 5/2$	2.50	2.27	1.34	128.6	0.00

^a Histidine beta carbons were kept fixed during geometry optimization, while all the other atoms are allowed to relax. ^b In the B3LYP calculations, the 6-31+G* and 6-31G* atomic basis sets are used for the superoxide radical anion and the remaining atoms, respectively.

functionals, whereas it appears smaller from the PBE calculations. The Hubbard U correction significantly improves on the geometry as well. The PBE calculations display a significant shortening of the distance between iron and proximal oxygen for all spin states. The systematically shorter Fe–proximal O distance in PBE calculations can be viewed as a consequence of the overhybridization of the electronic spin orbitals in this DFT scheme, which is essentially corrected by the Hubbard U interaction term. Indeed, the latter seems to overcorrect for the small Fe–O distance from PBE, but this point cannot be easily addressed on the basis of the available experimental data, because of the uncertainty in the attribution to the system with or without an additional proton. Nevertheless, the results of this section indicate that the adopted Hubbard U correction brings general agreement with the results from various high-quality hybrid functionals. However, the comparison of structural and spin-state energetics data does not provide a strong case for the use of the Hubbard U correction. In fact, the advantage of adopting the Hubbard U correction becomes apparent when we discuss the spin density analysis and the possible reaction products later in this article.

Table 3 reports the results for a larger model of the SOR active site, where the beta carbons of the histidine ligands were added, suitably saturated, and kept fixed during geometry optimization. All other atoms were allowed to relax. The comparison with the data in Table 2, especially at the PBE + U computational level, essentially corroborates the use of the small model system in all of the following calculations. For the larger model, the B3LYP calculation slightly favors the intermediate-spin state, in agreement with the PBE + U results. This stresses the facts that the high- and intermediate-spin states are essentially degenerate and that their energy difference is, indeed, on the order of the relative changes that can be induced by the choice of the atomic model.

Spin Density Analysis of the SOR–Dioxygen Complex in the Relevant Spin States. Table 4 reports the atomic Löwdin spin population of iron, proximal and distal oxygens, and trans thiolate sulfur in the complex intermediate- and high-spin states, at three levels of DFT approximation using the minimal active-site model. For the intermediate-spin state, pure DFT using the PBE XC functional (first row in the table) substantially underestimates the Löwdin spin densities around the iron atom and the two oxygens compared to the B3LYP

Table 4. Spin Density in the (Small) Model of the SOR Active Site for the Intermediate- and High-Spin States^a

method	q_{Fe}	$q_{\text{proximal O}}$	$q_{\text{distal O}}$	q_{O_2}	q_{S}
Intermediate Spin State					
PBE	2.94	0.02	-0.09	0.07	0.03
PBE + U	3.89	-0.43	-0.54	-0.97	0.02
B3LYP	3.72	-0.39	-0.53	-0.92	0.07
High Spin State					
PBE	3.76	0.49	0.40	0.89	0.12
PBE + U	3.88	0.48	0.54	1.02	0.02
B3LYP	3.77	0.49	0.51	1.00	0.06

^aLöwdin spin-up minus spin-down populations (q) on iron, oxygens, and sulfur are reported.

calculation. Such a discrepancy can also be visualized in Figure 4 and corresponds to a migration of spin-up electron charge toward the dioxygen. On the contrary, the spin density from the PBE + U approach shows excellent agreement with that from B3LYP, thus further supporting the supposition that the Hubbard U electron–electron interaction corrects for orbital overhybridization.

The three computational schemes yield similar atomic Löwdin spin populations for the high-spin complex (Figure 5), although they correspond to different spatial distributions. In fact, the net spin density on the dioxygen moiety has a σ^* character according to the PBE results and a π^* character according to the PBE + U and B3LYP calculations. Moreover, the two spin states differ only in the relative spin directions on iron and dioxygen (cf. Figures 4 and 5), in agreement with the discussion in the previous section.

According to both the PBE + U and B3LYP approaches, the dioxygen carries a net spin-up minus spin-down charge of about +1 and -1 in the high-spin and intermediate-spin state, respectively. The change in the sign of the electron spin density corresponds to the electron spin flipping

described above. The absolute value of the spin population, which is around unity in both the spin states, indicates that the superoxide is not reduced to the peroxide species (O_2^{2-}) as a mere consequence of its binding to the protein active site. This is also confirmed by the O–O bond length reported in Tables 2 and 3 for different computational levels. As shown below, the reduction process is likely to occur at a later stage of the enzymatic cycle, where it is assisted by proton transfer.

Peroxide Intermediate after the First Protonation. The protonation steps play an essential role in superoxide reduction. According to Nivière et al.,⁵⁴ the second intermediate in the enzyme reaction cycle is an iron–hydroperoxo complex, which is formed after one proton transfer to the dioxygen ligand. We studied the intermediate by means of DFT calculations on the same active site considered above with the addition of a H^+ ion. Structure optimizations were performed for the $S = 3/2$ and $S = 5/2$ states through plane-wave calculations at the PBE and PBE + U levels and atomic-basis-set calculations using the B3LYP hybrid functional. We obtained the ab initio Hubbard U values of 7.6 and 6.7 eV for the iron–hydroperoxo complex with high and intermediate spin, respectively. Thus, we assumed the same average U value of 7.2 eV for the comparison of the spin-state energies. Only oxygens and hydrogens were allowed to move relative to the crystal structure during geometry relaxation. As for the nonprotonated system, the inclusion of the imidazole beta carbons and the relaxation of the other heavy atoms do not significantly affect any of the main conclusions concerning structure and energetics.

In agreement with previous DFT studies,^{21,28} we consistently find, at different computational levels, that the first protonation occurs at the distal oxygen, irrespective of the

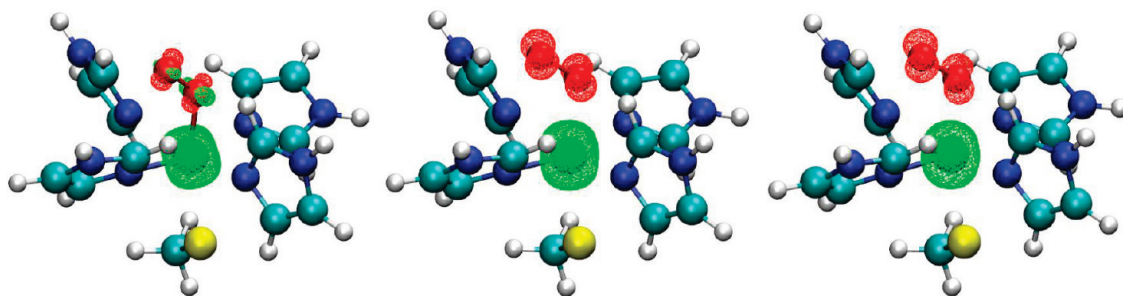


Figure 4. Spin density contours for the iron–dioxygen complex in the intermediate-spin state, from (left) PBE, (middle) PBE + U , and (right) B3LYP calculations. The positive and negative isovalues are shown in green and red, respectively.

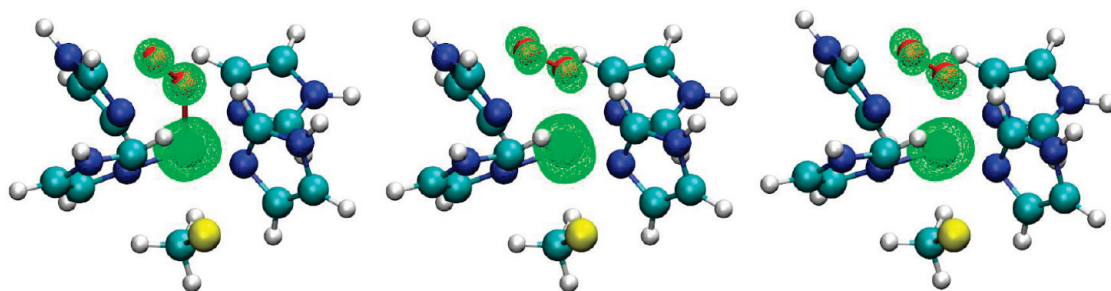


Figure 5. Spin density contours for the iron–dioxygen complex in the high-spin state, from (left) PBE, (center) PBE + U , and (right) B3LYP calculations. The positive spin population on the dioxygen is displayed by the isocontour in green.

Table 5. Structural Parameters and Relative Energies for the Model Iron–Hydroperoxo Intermediate in Different Spin States, from DFT Calculations at the PBE, PBE + *U*, and B3LYP Levels

method	spin state	$d_{\text{Fe-proximal O}}$ (Å)	$d_{\text{O-O}}$ (Å)	$\angle_{\text{Fe-O-O}}$ (deg)	energy (kcal/mol)
PBE	$S = 3/2$	1.67	1.91	129.8	4.11
	$S = 5/2$	1.99	1.47	124.8	0.00
PBE + <i>U</i>	$S = 3/2$	2.18	1.44	122.2	0.60
	$S = 5/2$	2.11	1.47	121.8	0.00
B3LYP	$S = 3/2$	1.91	1.47	123.9	8.00
	$S = 5/2$	2.00	1.47	123.9	0.00
expt ²¹		2.00		126	

Table 6. Löwdin Spin-up Minus Spin-down Population on Iron, Oxygens, and Trans Thiolate Sulfur in the SOR Active Site for the Intermediate and High Spin States of the Iron–Peroxo Complex

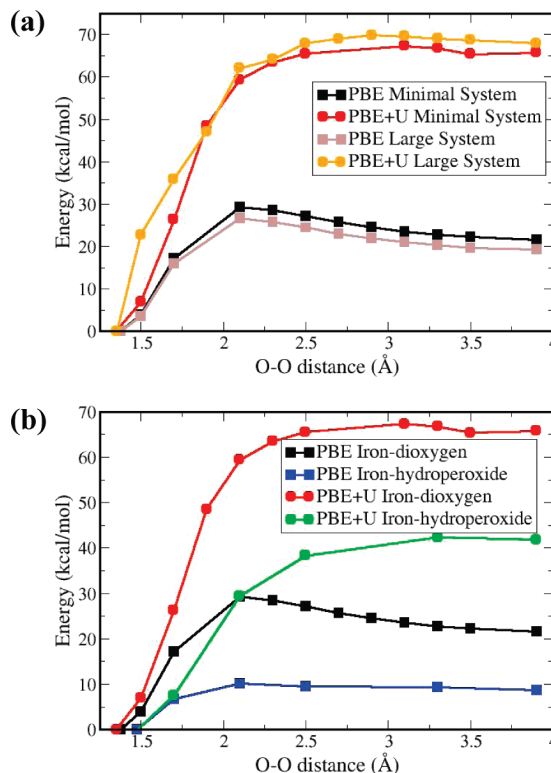
method	q_{Fe}	$q_{\text{proximal O}}$	$q_{\text{distal O}}$	q_{O_2}	q_{S}
Intermediate Spin State					
PBE	3.03	0.21	-0.41	-0.19	-0.04
PBE + <i>U</i>	3.82	-0.31	-0.11	-0.42	-0.36
B3LYP	3.18	0.03	0.00	0.03	-0.29
High Spin State					
PBE	3.97	0.29	0.05	0.34	0.36
PBE + <i>U</i>	4.24	0.22	0.03	0.25	0.27
B3LYP	4.16	0.20	0.03	0.23	0.31

active-site spin state. In fact, proton binding to the proximal oxygen is disfavored by steric hindrance.

Relevant structural parameters and energetics of the iron–hydroperoxo species in SOR, with one additional proton bonded to the distal oxygen, are reported in Table 5. The different computational setups consistently yield a high-spin complex lower in energy than the intermediate-spin one. Note that the O–O bond is broken in the intermediate-spin state from the PBE calculation. Moreover, the O–O bond length has a value typical of peroxide species, which points to its reduction after the first protonation. Therefore, the intermediate-spin and high-spin states mainly differ for the configuration of the electron spins on Fe, rather than for an electron flipping on the dioxygen, which explains the breaking of the quasidegeneracy between the two spin states.

The significant, yet not complete, reduction of the dioxygen after H^+ addition is confirmed by the spin population analysis in Table 6, where the spin-up minus spin-down charge on the peroxide (or, equivalently, on the dioxygen moiety, given that the spin density on the attached hydrogen is negligible) is significantly smaller than the unit electronic charge for all of the XC functionals employed. For both spin states, the PBE data display poor agreement with the B3LYP values. By assuming the B3LYP results as reference values, the adopted Hubbard *U* correction appears not to be effective for the system in the (excited) intermediate-spin state, where the disagreement is rather magnified. On the other hand, the Hubbard *U* correction determines a significant improvement for the high-spin state, which is the relevant one in the SOR catalytic reaction, according to the recent literature.^{21,52}

The comparison between Tables 4 and 6 highlights a donation of spin-down (spin-up) electron charge to the peroxide species from the iron and, through it, from the sulfur

**Figure 6.** Energy profiles for O–O cleavage in the (a) iron–dioxygen and (b) iron–hydroperoxide complexes in the high-spin state, obtained from PBE and PBE + *U* calculations on the two indicated SOR models.

ligand in the high-spin (intermediate-spin) state. In fact, the spin-up minus spin-down populations on iron and sulfur are increased (decreased) and the spin population on the dioxygen moiety is correspondingly decreased (increased), whereas no appreciable changes were obtained in the spin populations of the other atoms. The contribution of the sulfur ligand is an expression of its role in the enzymatic mechanism, which deserves further investigation. The direct contribution from iron indicated by both the PBE + *U* and B3LYP results is larger than that shown by the PBE values.

O–O Bond Cleavage in the Iron–(Hydro)Peroxide Complexes. Unlike heme-based peroxidases and many oxygenases, which promote the cleavage of the oxygen–oxygen bond to form reactive iron–oxo species,^{55–57} SORs form a stable iron–dioxygen intermediate that is subsequently involved in two protonation steps along the catalytic cycle.^{21,25–27} For the system under study, this behavior is well described by the theoretical curves in Figure 6. They were obtained at the PBE and PBE + *U* computational levels and represent the energy profiles of oxygen–oxygen homolysis in the high-spin state of the iron–dioxygen complex, for the two differently sized models of the SOR active site described above. In the minimal model, only oxygens and hydrogens are allowed to move during the geometry optimization at each fixed O–O bond distance, whereas in the larger model, the imidazole beta carbons are kept fixed and the other atoms are relaxed. Figure 6a displays the stability of the iron–dioxygen complex against the cleavage of the O–O bond. The energy barrier for the homolysis obtained by means of the PBE functional is 26.8 kcal/mol. This value

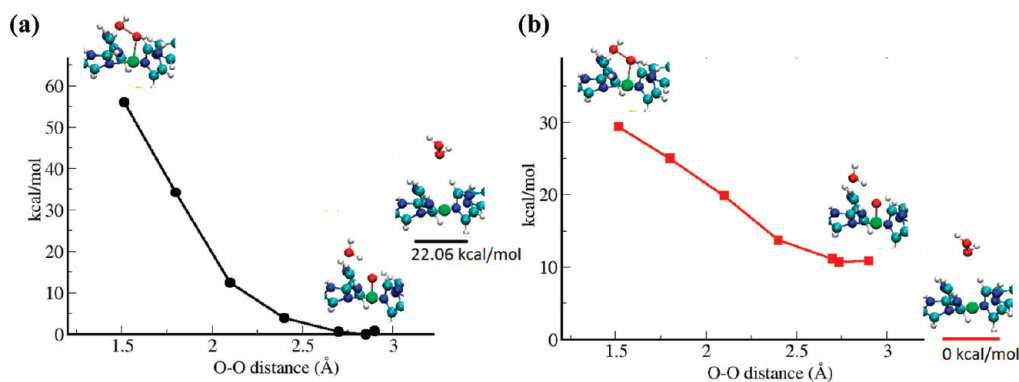


Figure 7. Energy profile for O–O bond cleavage when the H^+ ions are initially bonded to the same oxygen atom, compared to the total energy of the separate active site and hydrogen peroxide, from (a) PBE and (b) PBE + U calculations.

is in good agreement with the value of 24 kcal/mol from a previous study using the B88P86 functional without the Hubbard U correction.⁵⁶ Our PBE + U calculations indicate a stronger stability against O–O bond cleavage, with an energy barrier of 67.8 kcal/mol. The stabilization achieved through the Hubbard U correction can be ascribed to a weakening of the Fe–O interaction after correcting for the PBE overhybridization and a consequent strengthening of the O–O bond. Note that the larger model, where all of the atoms except for the imidazole beta carbons are relaxed, leads to quantitatively similar energy profiles, therefore showing that the relaxation of the ligands does not affect the energetics of the homolysis.

According to previous studies,⁵⁴ the first proton transfer occurs about 100 μs after the formation of the iron–dioxygen complex. It is also suggested,²⁸ and obtained by the present theoretical study, that the distal oxygen is the accepting site for the first protonation. Furthermore, the resulting iron–hydroperoxo complex has to be stable against peroxide cleavage, in order to allow the second protonation step to proceed, leading to hydrogen peroxide formation. This stability was investigated by evaluating the energy profile for the cleavage of the O–O bond in the iron–hydroperoxide complex, as shown in Figure 6b. Calculations were again carried out for the high-spin state. In agreement with the known biology at the SOR active site, our results suggest that the iron–hydroperoxide intermediate is stable after the first protonation process. The added H^+ ion weakens the stability of the molecular oxygen in the SOR active site (compare the blue and green curves with the black and red ones, respectively, in Figure 6b), but the hydroperoxo species is still clearly stable against cleavage. Note that, also in this case, the use of the Hubbard U correction to the PBE computational scheme yields a significant enhancement in the energy barrier against the cleavage of the O–O bond. This can entail a better description of SOR activity through the PBE + U computational approach, given that it matches with the behavior expected from this enzyme.

Second Protonation and Formation of Hydrogen Peroxide. As seen in the previous section, both reaction intermediates are stable against cleavage of the O–O bond. Accordingly, we study the reaction products after two protonations. Two scenarios can be envisaged: (i) In the first case, the second H^+ ion binds to the distal oxygen. This is

expected to lead to the cleavage of the O–O bond and thus to the formation of the iron–oxo reactive complex and a water molecule, similarly to what happens in cytochrome P450. (ii) Alternatively, the second H^+ ion binds to the proximal oxygen, which induces the full reduction of the superoxide to peroxide. Afterward, the hydrogen peroxide leaves the oxidized active site. In fact, the latter is the scenario consistent with the SOR mechanism, in line with experimental evidence.

To test the ability of our computational approach to discriminate between the above scenarios, we performed structural optimizations with the second proton initially attached to either the proximal oxygen or the distal one. The system is kept in the high-spin state, which is shown to be the most stable spin state after both the first and the second protonation, though being almost degenerate with the intermediate-spin state for the iron–dioxygen complex. The same Hubbard U value of 7.2 eV used for the iron–hydroperoxide complex is kept in this study. Indeed, for the oxidized active site and the iron–oxide complex, we obtained self-consistent U values of 7.4 and 7.6 eV, respectively. Therefore, $U = 7.2$ eV, which is quite close to these values, appears to be a convenient choice to study the mechanism of the whole catalytic cycle. The O–O bond breaks, forming a water molecule and an iron–oxide complex if the two protons are both bonded to the distal oxygen. Instead, when the two protons are initially bound to different oxygens, the products are always hydrogen peroxide and the oxidized active site. On the other hand, these final products are lower or higher in energy than the product pair made of the water molecule and the iron–oxide complex according to whether the Hubbard U correction is used or not, respectively. This is illustrated in Figure 7, where we report the energy profile for the O–O bond cleavage when the two H^+ ions are initially bonded to the same oxygen atom, so that oxygen separation is definitely observed along with the total energies of the separate active site and hydrogen peroxide species. Without Hubbard U correction, the water and iron–oxo products are more stable by 22.06 kcal/mol than the hydrogen peroxide and oxidized active-site products. On the contrary, with the Hubbard U correction, the system made of separate hydrogen peroxide and oxidized active site is more stable by 10.72 kcal/mol. Therefore, by introducing the Hubbard U interaction term we can establish the most stable product

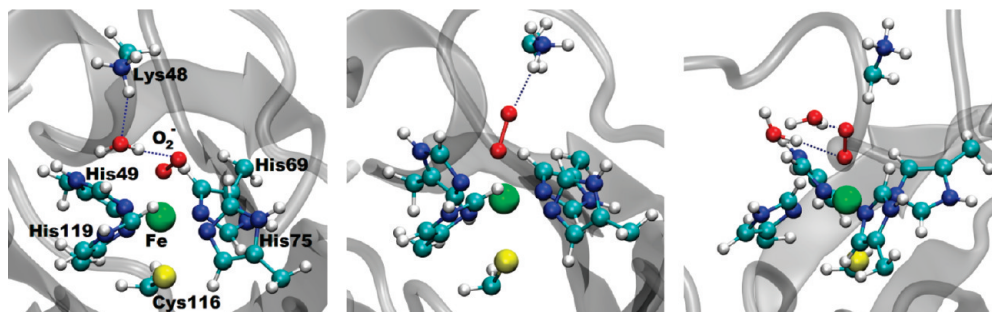


Figure 8. SOR active site in scenarios 1 (left), 2 (middle), and 3 (right).

pair and thus identify the most probable reaction path (i.e., the path starting from single protonations at both the oxygens, which is the only one leading to the expected product pair).

Possible Scenarios for the Proton-Transfer Steps. Our investigation of the possible scenarios for proton transfer to the dioxygen moiety starts with classical MD calculations, using the NAMD simulation package.⁵⁸ The initial configuration was again taken from the experimental X-ray structure¹ (PDB ID 2JI3), but the protein was mutated back to its wild-type form. The AMBER force field (ff03)⁵⁹ was adopted for all of the standard residues, whereas Bloechl charges⁶⁰ were used for the superoxide, Fe²⁺ ion, and its ligands. Harmonic potentials were applied to the metal-ion–ligand coordination. The details of the simulation are discussed in the Supporting Information. The force constants were obtained from our PBE + *U* calculations on the model active site. The atomic charges and the parameters are also listed in the Supporting Information. The rmsd value along the classical MD simulation, compared to the X-ray structure, is 1.8 Å for the entire-protein heavy atoms and 0.5 Å for the active-site residues. This shows that the overall structure is maintained well with the choice of the force field. The Lys48 residue is relatively flexible and the distance between the N ζ atom in Lys48 and Fe²⁺ varies from 4.8 to 11.3 Å, with an average distance of 7.76 Å. On the other hand, Glu114 is held more rigidly in the protein matrix, and the distance separating the C δ atom on the carboxylate group of Glu114 and Fe²⁺ fluctuates between 6.6 and 11.0 Å, with an average value of 8.8 Å.

We explore the corresponding hydrogen-bond networks that can convey proton transfer. A hydrogen bond is defined by the distance and the orientation of hydrogen-bond donor (H–Y) and acceptor (X), with the distance between X–Y smaller than 4.0 Å, and the X \cdots H–Y angle larger than 150°. In scenario 1, the dioxygen is hydrogen-bonded to a single water molecule that, in turn, is hydrogen-bonded to the (protonated) Lys⁴⁸, whereas in scenario 2, the Lys⁴⁸ residue is hydrogen-bonded directly to the superoxide. In scenario 3, only water molecules are hydrogen-bonded to superoxide. Lys⁴⁸ does not form any hydrogen bond with the first water shell of superoxide. The occurrence probabilities for scenarios 1–3 in the classical MD simulation are 11%, 3%, and 34%, respectively. In 52% of the time, the superoxide is not hydrogen-bonded to water or any side chains. Because the formation of hydrogen bonds is the prerequisite for proton transfer, we believe that the most populated scenario in which O₂[−] does not form any hydrogen bonds with other species

should not be significant in proton transfer. On the other hand, the average distances between the N ζ atom in Lys48 and Fe²⁺ are 7.68, 6.79, and 7.47 Å in scenarios 1–3, respectively. Whereas Lys48 comes closer to the metal center to form a direct hydrogen bond to superoxide in scenario 2, Lys48 stays more or less the same distance from the metal center in both scenarios 1 and 3, and the existence of an interstitial water molecule distinguishes the two scenarios.

Quantum chemical studies on the three scenarios were then performed. The model systems for the three representative snapshots from the classical MD represented in Figure 8 were obtained by replacing Cys¹¹⁶ with a methyl thiolate, Lys⁴⁸ with a methyl ammonium, and the four histidines with four neutral imidazole ligands. Only the oxygen and hydrogen atoms were allowed to move during the geometry optimizations. According to our results, in scenarios 1 and 2, proton transfer occurs spontaneously from Lys⁴⁸ to superoxide, either through a water molecule (scenario 1) or directly (scenario 2) in the PBE + *U* calculations, and there are no barriers in both of these proton transfers. In scenario 3, methyl ammonium was not included in the simulation, because it is not involved in the hydrogen-bond network with the superoxide. We found no proton transfer from the water molecules to the dioxygen during the geometry optimization. In this scenario, Lys⁴⁸ can still be connected to the superoxide through a longer hydrogen-bond chain. Because, in this case, the overall proton-transfer process involves successive hopping steps, the rate of transfer will diminish as the length of hydrogen-bond chain increases. In any case, the quantum chemical calculation for scenario 3 shows that proton transfer from water to superoxide forming a bare hydroxide ion is energetically unfavorable. This, however, does not rule out water as a possible proton source, because the hydroxide ion formed after protonation can be stabilized in the presence of other water molecules from the aqueous environment. B3LYP and PBE0 calculations were also performed at the geometries corresponding to scenarios 1 and 2. They show a proton transfer from Lys⁴⁸ to the superoxide, thus implying that the p*K*_a of the dioxygen moiety bonded to the SOR active site is higher than that of Lys⁴⁸. Although Lys⁴⁸ is believed to noticeably influence the catalytic activity of SOR, whether its side chain is directly involved in the proton-transfer process is still under debate.^{28,54,61} Moreover, some studies⁵⁴ of the dependence of the protonation rate on the pH suggest that water molecules are mainly involved in the first protonation step. In fact, the first protonation still occurs after K48I mutation, although at a slightly lower rate. This finding

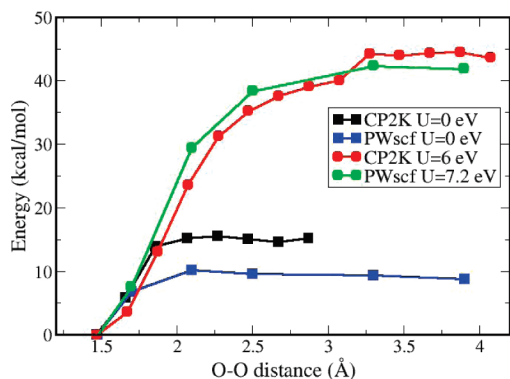


Figure 9. Comparison of the energy profiles for O–O bond cleavage in the iron–hydroperoxide complex, in its high-spin state, obtained using the PWscf program ($U = 0$ and 7.2 eV) and the CP2K code ($U = 0$ and 6 eV).

is partially consistent with our scenario 1, despite the fact that, in our simulations, the ultimate proton source is Lys⁴⁸.

Because scenarios 1 and 2 both occur readily in the classical MD trajectory spanning 60 ns, proton transfer should occur on the nanosecond time scale after the SOR–dioxygen intermediate is formed. However, pulse radiolysis experiments⁵⁴ provide a proton-transfer rate of $(4280 \pm 300) \text{ s}^{-1}$ at neutral pH. Such a huge discrepancy can arise for several reasons. For example, the X-ray structure from which we start our classical MD simulation might be already in the free energy basin that favors the proton transfer. Moreover, our classical MD and subsequent quantum-chemical simulations might have missed some key elements that deter quick proton transfer.

The proton-transfer mechanism was investigated further using the QM/MM approach,¹⁶ which has been demonstrated to be an excellent tool for studying the reactivity of biological systems.^{62,63} The simulation details were provided in the previous section. The high-spin state was chosen for the whole simulation. The occupation matrix used for the Hubbard U implementation in CP2K was obtained by projecting the Kohn–Sham orbitals on a Gaussian basis of d symmetry. Then, the Hubbard U value was chosen in such a way that the structure and energetics of the active site models from CP2K agreed well with the PBE + U results from PWscf. In particular, the energy profiles for the O–O bond cleavage in the iron–hydroperoxo complex from CP2K and PWscf calculations are reported in Figure 9. We find that a U value of 6.0 eV yields an excellent agreement in the energy profile. Furthermore, the spin density obtained with this U value in CP2K (cf. Figures 10 and 5) is essentially the same as that obtained from the PBE + U calculation in PWscf. Specifically, the π -like spin distribution on the dioxygen moiety is correctly reproduced. Moreover, the adopted U value leads to the correct reaction products. In fact, the product pair made of iron–oxo radical and water is more stable than that consisting of oxidized SOR and hydrogen peroxide by 20.51 kcal/mol at the PBE level and is stabilized by 6.93 kcal/mol using the choice of $U = 6$ eV.

The initial configuration for the QM/MM simulation was taken from a snapshot corresponding to scenario 1 in Figure 7. As mentioned above, there is a hydrogen-bond chain

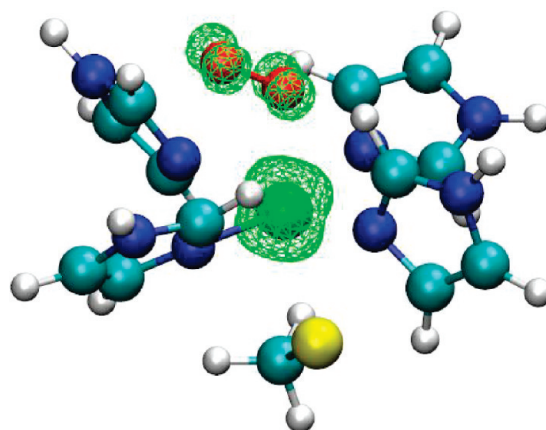


Figure 10. Spin density of the iron–dioxygen complex calculated with PBE + U in CP2K. The positive spin population on the dioxygen is displayed by the isocontour in green.

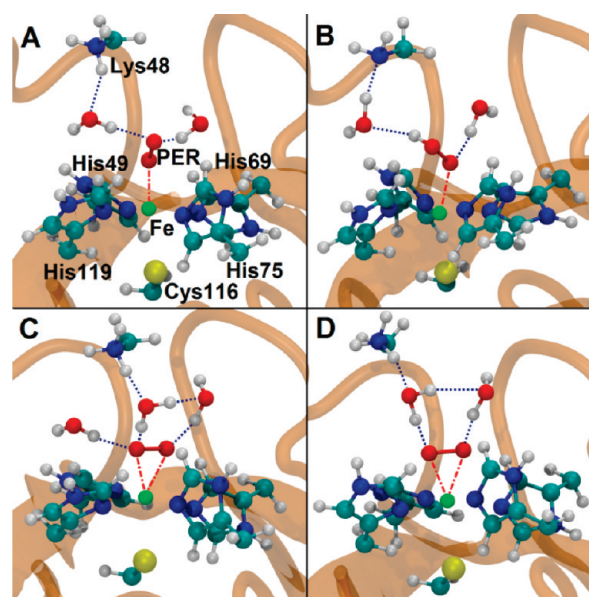


Figure 11. QM/MM simulation starting from (A) a snapshot after 10 ns of classical MD simulation, which is representative of scenario 1. (B) Snapshot after 1 ps of QM/MM simulation showing the proton temporarily transferred to the superoxide. (C) Snapshot taken after 3.5 ps displaying a side-on conformation of the superoxide, which is stably attained and preserved until the end of the simulation (D), at 8.5 ps.

linking Lys⁴⁸, a water molecular, and the O_2^- moiety. We used $U = 6$ eV throughout the MD run and a total simulation time of 8.5 ps. In the first 1.8 ps of simulation, the protons in the hydrogen-bonded chain oscillate between Lys⁴⁸, water, and O_2^- . Parts A and B of Figure 11 show two representative snapshots at 0 and 1 ps, respectively. However, the O_2^- moiety never permanently captures the proton from the hydrogen-bond chain. Instead, after about 1.8 ps, it adopts a side-on conformation, as shown in Figure 11C. There is no proton oscillation along the hydrogen-bond chain after the side-on conformation is attained, and protonation does not occur for the rest of the MD simulation (see Figure 11D).

Whereas our quantum calculations on the model active site from the crystal structure²¹ in the various spin states always give a stable conformation of the end-on type, our

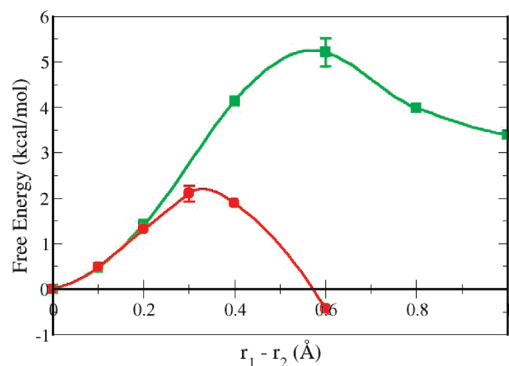


Figure 12. Free energy profile from side-on to end-on conformation from CP2K calculations with a Hubbard U value of 6.0 eV. The conformational rearrangement corresponding to the red curve is accompanied by a proton transfer from Lys⁴⁸ to the superoxide. The green curve was obtained by restraining the H⁺ on Lys⁴⁸, in order to isolate the conformational change from the protonation step.

QM/MM study stabilizes the side-on configuration, as also suggested by the studies on the natural enzyme^{27,52} and biometric models.^{64,65} According to our QM/MM results, the side-on dioxygen conformer hinders fast proton transfer. We performed thermodynamic integrations⁶⁶ by constraining $r_1 - r_2$, where r_1 and r_2 are the Fe–O_{distal} and Fe–O_{proximal} distances, respectively. Each window was simulated for at least 3 ps after equilibration was attained. We thus obtained the free energy profile connecting the side-on and end-on conformations. As shown in Figure 12 (red curve), we obtained a barrier of ~ 2.3 kcal/mol in going from the side-on to the end-on complex. When the O–O moiety is pushed end-on in the free-energy calculation, the proton transfer occurs immediately, suggesting a barrierless transition. The value for the height of the energy barrier is of the same order of magnitude as its estimate from the experimental proton-transfer reaction rate⁵⁴ using the transition-state theory, which is about 10 kcal/mol. The actual comparison with experiment needs to consider that the proton donation from Lys⁴⁸, which accompanies the conformational rearrangement in our QM/MM simulation, can be hindered by displacements of the flexible loop bearing Lys⁴⁸. The energy profile for the transition from the side-on to the end-on conformation without proton transfer is represented by the green curve in Figure 12, which was obtained by making the proton sources, that is, Lys⁴⁸ and the surrounding water molecules, rigid.

To understand why QM/MM suggests a different stable conformation than that obtained from our QM calculations on the model active site from the SOR crystal structure, we constructed the model active site from the system structure after 3.5 ps of QM/MM simulation (see Figure 11C), by including the Lys⁴⁸ residue and the relevant water molecules around the active site. In this snapshot, the side-on O₂[−] is hydrogen-bonded to the His imidazoles and to the water molecules, which, in turn, are hydrogen-bonded to Lys⁴⁸. The orientations of the His ligands differ from those in the crystal structure and appear to stabilize the side-on dioxygen conformer. This stabilization does not correspond to a significant increase of the Fe–imidazole distances, as claimed in a recent theoretical study.²⁸ In fact, we calculated the

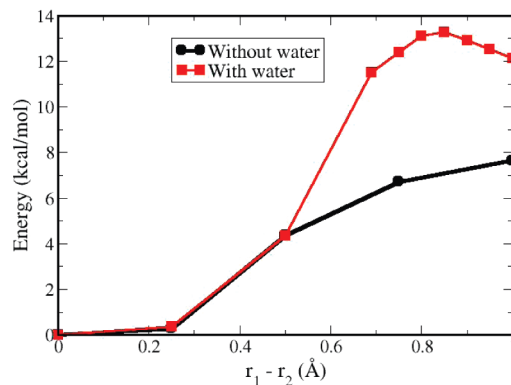


Figure 13. Energy of the end-on conformer, relative to the side-on conformer, against the $r_1 - r_2$ internal coordinate, with (red) and without (black) interstitial water.

average distance between Fe and the four His ligand nitrogen atoms from our QM/MM simulation and found only a 0.03 Å increase compared to the experimental structure.²¹ Note that fixing the heavy atoms to the experimental structure in our gas-phase calculations is not the reason why the side-on conformation was not observed from previous calculations^{21,28} and our calculations involving a larger model system also gave an end-on conformation. Important further stabilization comes from the hydrogen bonds with the surrounding water molecules. The energy of the end-on conformer relative to the side-on conformer is shown in Figure 13 as a function of the $r_1 - r_2$ internal coordinate, which measures the departure from the side-on conformation. The two energy profiles, which were obtained from PBE + U calculations with CP2K on the model active site taken from one QM/MM snapshot, display the stability of the side-on conformer relative to the end-on one. Their difference gives a measure of the contribution to side-on stabilization by waters, even if an exact superposition of the hydrogen bonds to histidines and water molecules is not assumed. Ultimately, our findings point to the stabilization of the side-on superoxide conformer, which is not amenable to protonation according to our calculations, in a significant percentage of the protein conformations as a major motif in justifying the observed rate for the first protonation step of the enzymatic cycle.

Conclusions

Our extensive quantum-mechanical study of the SOR active site through various DFT approaches leads to the following conclusions: (i) In a vacuum, keeping the coordinates of the heavy atoms as in the crystal structure, the end-on conformation of the dioxygen ligand is favored to the side-on one before and after the first protonation step. In general, the dioxygen conformation can depend on the conformation of the active site. (ii) For the iron–dioxygen complex, the quasidegeneracy of the high-spin and intermediate-spin states is rationalized in terms of exchange forces between the dioxygen ligand and the central iron. The first protonation stabilizes the high-spin state and induces a significant reduction of the superoxide. (iii) The main features of the active-site structure and energetics are captured and shown to be robust against different hybrid-DFT schemes, the

standard DFT-PBE approach, and its Hubbard U correction. (iv) The observed electron charge donation from the sulfur ligand to the dioxygen moiety through Fe suggests a specific role for the sulfur ligand that is worthy of further investigation. (v) The adopted Hubbard U correction to the standard DFT-PBE approximation leads to results in general agreement with the high-level hybrid-DFT computational schemes using the B3LYP and PBE0 XC functionals. Moreover, PBE wrongly predicts the reaction products after two protonations, whereas the Hubbard U term corrects for this deficiency. The present results should encourage the use of the correction term in QM/MM calculations, thus allowing the study of the active-site region without additional computational cost as compared with standard DFT.

The use of classical MD allows exploration of possible scenarios for the protonation steps involved in the enzymatic cycle. They are essentially characterized by hydrogen-bond networks that can convey protons toward the dioxygen species. Discrimination among the possible scenarios, as well as their effectiveness in the protonation steps, is gained through a QM/MM approach. The latter provides with the following indications on the SOR mechanism: (i) The Lys⁴⁸ residue is one of the proton sources for the first protonation, either directly or through the interstitial water. (ii) The imidazole rings of the His ligands and the interstitial water afford a pattern of interactions and hydrogen bonds that appears to play an essential role in the enzyme catalytic mechanism. In fact, whereas the hydrogen-bond pathways involving the water molecules could convey protons to the iron–dioxygen complex, the stabilization of the dioxygen moiety in a side-on conformation by the same hydrogen bonds and especially by the interactions with the histidine ligands prevents an immediate proton transfer. This conclusion is, indeed, irrespective of the presence of possible proton sources other than Lys⁴⁸. (iii) We find an energy barrier between the side-on and end-on dioxygen conformations. This barrier is of the correct order of magnitude, as compared with recent experimental findings.

Future work will involve the QM/MM study of scenario 2 and the investigation of whether the active site can afford a side-on dioxygen conformation without the presence of interstitial water molecules. Moreover, the detailed structural change in the active site in response to the formation of the side-on conformer can be fruitfully studied. In particular, it has been shown that the stability of the side-on conformation is related to the Fe–S ligand bond strength.²⁷ Furthermore, wild-type and mutated enzymes show different yields in the side-on conformation. Finally, the first protonation still occurs after the K48I mutation, although at a slightly lower rate, which points to the presence of another proton source in addition to Lys⁴⁸ that needs to be investigated.

The methodology presented and widely tested in this work appears to be a valid tool for the above-mentioned future explorations. Moreover, our present results concerning the chemistry at the SOR active site, and especially the hindrance to the first protonation provided by the side-on dioxygen conformation, offer a contribution to the understanding of the SOR catalytic cycle. In fact, our work might aid in suggesting new mutations useful to control the enzymatic

mechanism. We believe that the Hubbard- U -corrected DFT scheme within the QM/MM approach can be a fruitful approach to handle transition-metal enzymatic systems.

Acknowledgment. We thank the National Institutes of Health for supporting this work under Grant GM067689.

Supporting Information Available: Details of our classical molecular dynamics, atomic charges used in the SOR active site obtained from DFT + U calculations (Table S1), and force field parameters also calculated from DFT + U calculations (Table S2). This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- Halliwell, B.; Gutteridge, J. *Free Radicals in Biology and Medicine*, 3rd ed.; Oxford University Press: Oxford, U.K., 1999; pp 31–32.
- Mccord, J. M.; Fridovic, I. *J. Biol. Chem.* 1969244 (22), 6049.
- Jenney, F. E.; Verhagen, M. F. J. M.; Cui, X. Y.; Adams, M. W. W. *Science* **1999**, 286 (5438), 306.
- Niviere, V.; Fontecave, M. *J. Biol. Inorg. Chem.* **2004**, 9 (2), 119.
- Brines, L. M.; Kovacs, J. A. *Eur. J. Inorg. Chem.* 2007 (1), 29.
- Maritim, A. C.; Sanders, R. A.; Watkins, J. B. *J. Biochem. Mol. Toxicol.* **2003**, 17 (1), 24.
- Ihara, Y.; Chuda, M.; Kuroda, S.; Hayabara, T. *J. Neurol. Sci.* **1999**, 170 (2), 90.
- Kocatürk, P. A.; Akbostanci, M. C.; Tan, F.; Kavas, G. *Ö. Pathophysiology* **2000**, 7 (1), 63.
- Marklund, S. L.; Adolfsson, R.; Gottfries, C. G.; Winblad, B. *J. Neurol. Sci.* **1985**, 67 (3), 319.
- De Leo, M. E.; Borrello, S.; Passantino, M.; Palazzotti, B.; Mordente, A.; Daniele, A.; Filippini, V.; Galeotti, T.; Masullo, C. *Neurosci. Lett.* **1998**, 250 (3), 173.
- Fortunato, G.; Pastinese, A.; Intrieri, M.; Lofrano, M. M.; Gaeta, G.; Censi, M. B.; Boccalatte, A.; Salvatore, F.; Sacchetti, L. *Clin. Biochem.* **1997**, 30 (7), 569.
- Burdon, R. H. *Free Radical Biol. Med.* **1995**, 18 (4), 775.
- Toh, Y.; Kuninaka, S.; Mori, M.; Oshiro, T.; Ikeda, Y.; Nakashima, H.; Baba, H.; Kohnoe, S.; Okamura, T.; Sugimachi, K. *Oncology* **2000**, 59 (3), 223.
- Shearer, J.; Nehring, J.; Lovell, S.; Kaminsky, W.; Kovacs, J. A. *Inorg. Chem.* **2001**, 40 (22), 5483.
- Warshel, A.; Levitt, M. *J. Mol. Biol.* **1976**, 103 (2), 227.
- Laio, A.; VandeVondele, J.; Rothlisberger, U. *J. Chem. Phys.* **2002**, 116 (16), 6941.
- Cococcioni, M.; Dal Corso, A.; de Gironcoli, S. *Phys. Rev. B* **2003**, 67 (9), 094106.
- Cococcioni, M.; de Gironcoli, S. *Phys. Rev. B* **2005**, 71 (3), 035105.
- Hubbard, J. *Proc. R. Soc. London Ser. A* 1963276 (Dec), 238.
- Nowak, M. J.; Lapinski, L.; Kwiatkowski, J. S.; Leszczynski, J. In *Molecular Structure and Infrared Spectra of the DNA Bases and Their Derivatives: Theory and Experiment*; Leszczynski, J., Ed.; World Scientific: Singapore, 1997; pp 140–216.

- (21) Katona, G.; Carpentier, P.; Niviere, V.; Amara, P.; Adam, V.; Ohana, J.; Tsanov, N.; Bourgeois, D. *Science* **2007**, *316* (5823), 449.
- (22) Adam, V.; Royant, A.; Niviere, V.; Molina-Heredia, F. P.; Bourgeois, D. *Structure* **2004**, *12* (9), 1729.
- (23) Yeh, A. P.; Hu, Y. L.; Jenney, F. E.; Adams, M. W. W.; Rees, D. C. *Biochemistry* **2000**, *39* (10), 2499.
- (24) Clay, M. D.; Jenney, F. E.; Hagedoorn, P. L.; George, G. N.; Adams, M. W. W.; Johnson, M. K. *J. Am. Chem. Soc.* **2002**, *124* (5), 788.
- (25) Bukowski, M. R.; Halfen, H. L.; van den Berg, T. A.; Halfen, J. A.; Que, L. *Angew. Chem., Int. Ed.* **2005**, *44* (4), 584.
- (26) Mathe, C.; Mattioli, T. A.; Horner, O.; Lombard, M.; Latour, J. M.; Fontecave, M.; Niviere, V. *J. Am. Chem. Soc.* **2002**, *124* (18), 4966.
- (27) Mathe, C.; Niviere, V.; Houee-Levin, C.; Mattioli, T. A. *Biophys. Chem.* **2006**, *119* (1), 38.
- (28) Silaghi-Dumitrescu, R.; Silaghi-Dumitrescu, L.; Coulter, E. D.; Kurtz, D. M. *Inorg. Chem.* **2003**, *42* (2), 446.
- (29) Costas, M.; Mehn, M. P.; Jensen, M. P.; Que, L. *Chem. Rev.* **2004**, *104* (2), 939.
- (30) Shearer, J.; Scarrow, R. C.; Kovacs, J. A. *J. Am. Chem. Soc.* **2002**, *124* (39), 11709.
- (31) Quantum-ESPRESSO is a community project for high-quality quantum-simulation software, based on density-functional theory, and coordinated by Paolo Giannozzi. See <http://www.quantum-espresso.org> and <http://www.pwscf.org>.
- (32) (a) Aprà, E.; Windus, T. L.; Straatsma, T. P.; Bylaska, E.; de Jong, W.; Hirata, S.; Valiev, M.; Hackler, M.; Pollack, L.; Kowalski, K.; Harrison, R.; Dupuis, M.; Smith, D. M. A.; Nieplocha, J.; Tipparaju, V.; Krishnan, M.; Auer, A. A.; Brown, E.; Cisneros, G.; Fann, G.; Fruchtl, H.; Garza, J.; Hirao, K.; Kendall, R.; Nichols, J.; Tsemekhman, K.; Wolinski, K.; Anchell, J.; Bernholdt, D.; Borowski, P.; Clark, T.; Clerc, D.; Dachsel, H.; Deegan, M.; Dylla, K.; Elwood, D.; Glendening, E.; Gutowski, M.; Hess, A.; Jaffe, J.; Johnson, B.; Ju, J.; Kobayashi, R.; Kutteh, R.; Lin, Z.; Littlefield, R.; Long, X.; Meng, B.; Nakajima, T.; Niu, S.; Rosing, M.; Sandrone, G.; Stave, M.; Taylor, H.; Thomas, G.; van Lenthe, J.; Wong, A.; Zhang, Z. *NWChem, A Computational Chemistry Package for Parallel Computers*, version 4.7; Pacific Northwest National Laboratory: Richland, WA, 2005.
- (b) Kendall, R. A.; Aprà, E.; Bernholdt, D. E.; Bylaska, E. J.; Dupuis, M.; Fann, G. I.; Harrison, R. J.; Ju, J.; Nichols, J. A.; Nieplocha, J.; Straatsma, T. P.; Windus, T. L.; Wong, A. T. *Comput. Phys. Commun.* **2000**, *128*, 260.
- (33) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77* (18), 3865.
- (34) Vanderbilt, D. *Phys. Rev. B* **1990**, *41* (11), 7892.
- (35) Cremer, D. *Mol. Phys.* **2001**, *99* (23), 1899.
- (36) Becke, A. D. *J. Chem. Phys.* **1993**, *98* (7), 5648.
- (37) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37* (2), 785.
- (38) Adamo, C.; Barone, V. *J. Chem. Phys.* **1999**, *110* (13), 6158.
- (39) Keal, T. W.; Tozer, D. J. *J. Chem. Phys.* **2005**, *123* (12), 121103.
- (40) Becke, A. D. *J. Chem. Phys.* **1993**, *98* (2), 1372.
- (41) Kulik, H. J.; Cococcioni, M.; Scherlis, D. A.; Marzari, N. *Phys. Rev. Lett.* **2006**, *97* (10), 103001.
- (42) Migliore, A.; Sit, P. H. L.; Klein, M. L. *J. Chem. Theory Comput.* **2009**, *5* (2), 307.
- (43) VandeVondele, J.; Krack, M.; Mohamed, F.; Parrinello, M.; Chassaing, T.; Hutter, J. *Comput. Phys. Commun.* **2005**, *167* (2), 103.
- (44) Goedecker, S.; Teter, M.; Hutter, J. *Phys. Rev. B* **1996**, *54* (3), 1703.
- (45) Hartwigsen, C.; Goedecker, S.; Hutter, J. *Phys. Rev. B* **1998**, *58* (7), 3641.
- (46) Laino, T.; Mohamed, F.; Laio, A.; Parrinello, M. *J. Chem. Theory Comput.* **2005**, *1* (6), 1176.
- (47) Nose, S. *J. Chem. Phys.* **1984**, *81* (1), 511.
- (48) Hoover, W. G. *Phys. Rev. A* **1985**, *31* (3), 1695.
- (49) Wang, D.; Thiel, W. *J. Mol. Struct. (THEOCHEM)* **2009**, *898* (1–3), 90.
- (50) Shaik, S.; Cohen, S.; Wang, Y.; Chen, H.; Kumar, D.; Thiel, W. *Chem. Rev.* **2009**, *110* (2), 949.
- (51) Koch, W.; Holthausen, M. C. *A Chemist's Guide to Density Functional Theory*; Wiley: New York, 2000.
- (52) Horner, O.; Mouesca, J. M.; Oddou, J. L.; Jeandey, C.; Niviere, V.; Mattioli, T. A.; Mathe, C.; Fontecave, M.; Maldivi, P.; Bonville, P.; Halfen, J. A.; Latour, J. M. *Biochemistry* **2004**, *43* (27), 8815.
- (53) Jensen, F. *Introduction to Computational Chemistry*, 2nd ed.; John Wiley and Sons Ltd.: Chichester, U.K., 2007; pp 153–159.
- (54) Niviere, V.; Asso, M.; Weill, C. O.; Lombard, M.; Guigliarelli, B.; Favaudon, V.; Houee-Levin, C. *Biochemistry* **2004**, *43* (3), 808.
- (55) Lehnert, N.; Neese, F.; Ho, R. Y. N.; Que, L.; Solomon, E. I. *J. Am. Chem. Soc.* **2002**, *124* (36), 10810.
- (56) Lehnert, N.; Ho, R. Y. N.; Que, L.; Solomon, E. I. *J. Am. Chem. Soc.* **2001**, *123* (51), 12802.
- (57) Decker, A.; Solomon, E. I. *Curr. Opin. Chem. Biol.* **2005**, *9* (2), 152.
- (58) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kale, L.; Schulten, K. *J. Comput. Chem.* **2005**, *26* (16), 1781.
- (59) Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J.; Kollman, P. *J. Comput. Chem.* **2003**, *24* (16), 1999.
- (60) Blochl, P. E. *J. Chem. Phys.* **1995**, *103* (17), 7422.
- (61) Emerson, J. P.; Coulter, E. D.; Cabelli, D. E.; Phillips, R. S.; Kurtz, D. M. *Biochemistry* **2002**, *41* (13), 4348.
- (62) De Vivo, M.; Dal Peraro, M.; Klein, M. L. *J. Am. Chem. Soc.* **2008**, *130* (33), 10955.
- (63) Ho, M. H.; Vivo, M. D.; Peraro, M. D.; Klein, M. L. *J. Chem. Theory Comput.* **2009**, *5* (6), 1657.
- (64) Roelfes, G.; Vrajmasu, V.; Chen, K.; Ho, R. Y. N.; Rohde, J. U.; Zondervan, C.; la Crois, R. M.; Schudde, E. P.; Lutz, M.; Spek, A. L.; Hage, R.; Feringa, B. L.; Munck, E.; Que, L. *Inorg. Chem.* **2003**, *42* (8), 2639.
- (65) Girerd, J. J.; Banse, F.; Simaan, A. Characterization and Properties of Non-Heme Iron Peroxo Complexes. In *Metal-Oxo and Metal-Peroxo Species in Catalytic Oxidations*; Springer: Berlin, 2000; pp 145–177.
- (66) Ciccotti, G.; Ferrario, M.; Hynes, J. T.; Kapral, R. *Chem. Phys.* **1989**, *129* (2), 241.

JCTC

Journal of Chemical Theory and Computation

Approaching Elastic Network Models to Molecular Dynamics Flexibility

Laura Orellana,^{†,‡} Manuel Rueda,^{†,§} Carles Ferrer-Costa,[†] José Ramón Lopez-Blanco,^{||} Pablo Chacón,^{||} and Modesto Orozco^{*,†,‡}

Joint Research Program in Computational Biology from the Institute for Research in Biomedicine Barcelona (IRBB) and Barcelona Supercomputing Center (BSC), Barcelona, Spain, Chemical and Physical Biology, Centro de Investigaciones Biológicas, Madrid, Spain, Departament de Bioquímica i Biologia Molecular, Universitat de Barcelona, Barcelona, Spain, and Skaggs School of Pharmacy, University of California—San Diego, La Jolla, California 92093

Received February 16, 2010

Abstract: Elastic network models (ENMs) are coarse-grained descriptions of proteins as networks of coupled harmonic oscillators. However, despite their widespread application to study collective movements, there is still no consensus parametrization for the ENMs. When compared to molecular dynamics (MD) flexibility in solution, the ENMs tend to disperse the important motions into multiple modes. We present here a new ENM, trained against a database of atomistic MD trajectories. The role of residue connectivity, the analytical form of the force constants, and the threshold for interactions were systematically explored. We found that contacts between the three nearest sequence neighbors are crucial determinants of the fundamental motions. We developed a new general potential function including both the sequential and spatial relationships between interacting residue pairs which is robust against size and fold variations. The proposed model provides a systematic improvement compared to standard ENMs: Not only do its results match the MD results—even for long time scales—but also the model is able to capture large X-ray conformational transitions as well as NMR ensemble diversity.

1. Introduction

Protein functions largely depend on the intrinsic flexibility of their structures; even processes such as ligand binding or catalysis, in which the overall shape or surface properties play a dominant role, are coupled to local movements of the polypeptide backbone.¹ The intrinsic deformability of different protein families seems to guide structural changes along evolution, and deformation patterns (i.e., the large-scale motions) are extremely conserved in proteins displaying a common function.² Unfortunately, despite promising ad-

vances,³ the experimental study of large-scale dynamics is still difficult, and a large amount of information comes from theoretical calculations. Among the computational approaches to tackle the question of protein flexibility, molecular dynamics (MD)^{4–6} is probably the most accurate, since it is based on a rigorous physical formalism and a thorough parametrization from quantum-mechanical and experimental measurements. Although the high computational cost still limits atomistic simulations to the nanosecond to microsecond time scale, the principal component analysis (PCA) of MD trajectories—also called the essential dynamics (ED)⁷ approach—provides valuable information on large-scale functional motions, as we will discuss below. An alternative to MD to reach biologically relevant time and length scales is coarse-grained (CG) models,⁸ which simplify both the protein representation and the potential functions. Among these methods, the elastic network models (ENMs) are

* Corresponding author phone: (+34)934037156; fax: (+34)934037157; e-mail: modesto.orozco@irbbarcelona.org.

[†] Joint Research Program in Computational Biology from the IRBB and BSC.

[‡] Universitat de Barcelona.

[§] University of California—San Diego.

^{||} Centro de Investigaciones Biológicas.

probably the most widely used ones.⁹ The ENM potential is defined by a network of springs connecting the C^α atoms in a topology matrix Γ (known as the Kirchhoff matrix) of inter-residue contacts, where the ij th element is equal to -1 if nodes (i.e., residues) i and j are within the cutoff distance r_c or 0 otherwise and the diagonal elements (ii th) are equal to residue connectivity:

$$\Gamma_{ij} = \begin{cases} -1 & \text{if } d_{ij} \leq r_c \\ 0 & \text{if } d_{ij} > r_c \end{cases} \quad \Gamma_{ii} = - \sum_{kk \neq i}^N \Gamma_{ik} \quad (1)$$

The topology of the C^α network may be alternatively expressed in terms of a stiffness matrix, whose elements are the Hookean force constants, K_{ij} , acting between any pair of nodes i, j :

$$K_{ij} = 0.5\xi\Gamma_{ij} \quad (2)$$

where ξ is a constant which may or may not have the same value for all pairs, depending on the model. Hence, the overall potential energy of the network is given by

$$E = \sum_{i \neq j} K_{ij} (R_{ij} - R_{ij}^0)^2 \quad (3)$$

where R_{ij} and R_{ij}^0 are the instantaneous and reference (equilibrium) distances between each pair of α -carbons i and j . The functional in eq 3 can be implemented into Monte Carlo or dynamics algorithms¹⁰ to obtain ensembles of accessible configurations or within the elastic network normal mode analysis approach (NMA)^{11,12} to build the Hessian matrix of the potential. Within the anisotropic network model (ANM)¹³ approach, diagonalization of the Hessian directly yields a set of eigenvectors and eigenvalues (in energy or frequency units) which together define the near-equilibrium harmonic deformability space. In spite of this extreme simplicity, the lowest frequency modes of the ENMs provide descriptions of large-scale flexibility in good agreement with empirical and theoretical data, being especially well-suited to trace cooperative domain and segment movements. However, it cannot be ignored that ENMs are based on a harmonic, near-equilibrium approach and a rigid topology and thus have problems in capturing large anharmonic motions that can (in principle) be traced by MD simulations. Furthermore, there is no consensus parametrization, and the diverse models are often fitted to each particular problem.

Many attempts have been made to improve the robustness and generality of ENMs, for example, developing methods where atoms are grouped into rigid blocks,¹⁴ scaling differently covalent and noncovalent contacts,¹⁵ or using Markovian approaches¹⁶ to define the coarse-graining. The use of an isotropic constant and a cutoff is appealing for its simplicity, but can lead to different outcomes depending on the selected threshold for interactions¹⁷ (see also the Results and Discussion). Therefore, to avoid the use of an arbitrary cutoff, the discrete Hamiltonian is sometimes replaced by continuous functions that scale the force constants with an inverse power of the inter-residue distance. For example, Hinsen et al.¹⁸ derived a function for the spring strength by fitting to a local minimum from a single 1.5 ns MD simulation of one protein. This force constant definition

proposed stronger couplings for backbone neighbors and a sixth power of distance for the rest of the interactions. The distinction of short- and long-range terms was, however, dependent on a short 4 Å cutoff, and the formulation also included a protein-fitted scaling factor for the global energetics. Kovacs et al.¹⁹ proposed a simpler sixth-power exponential which did not require any cutoff or scaling factor. Other authors have also used several distance-dependent force constants,²⁰ including sometimes specific short-range terms (see ref 21, also based on MD) or, alternatively, bond cutoffs for chain neighbors.²² Recently, Jernigan et al.²³ suggested an inverse-square function for the reproduction of B factors, but they found that the resulting stronger long-range cohesion prevented discrete domains from moving properly.

Attempts to refine and improve ENMs have a common drawback: the lack of reliable experimental data on protein flexibility in solution, mostly coming from nuclear magnetic resonance (NMR) spectroscopy relaxation measurements²⁴ and to a lesser extent neutron scattering data,²⁵ both available for very few proteins. Therefore, in most studies so far, ENMs have been validated by fitting the calculated atomic fluctuations to B factors found in the crystal, in some cases to the degree of reaching an almost perfect fit.²⁶ Nevertheless, the use of X-ray B factors as a reference for flexibility in solution has been highly controversial,^{27–29} since they are subject to crystal packing effects, among other biases such as internal static disorder or refinement errors.³⁰ Other indirect sources of flexibility data for calibration and benchmarking have been the study of the environment-dependent conformational space of proteins^{23,31} and, more recently, the analysis of NMR ensembles,^{32,33} including comparisons with their RMSDs.³⁴ However, in the first case no guarantee exists that conformational changes induced, for example, by the presence of other molecules match the intrinsic deformation pattern of apo proteins. Furthermore, principal components predicted from PCA of selected NMR ensembles agree with the ED modes,³⁵ but caution must be taken since local diversity of NMR structures may also be a sign of experimental uncertainty due to missing data. In summary, there is a dramatic lack of direct experimental information on protein flexibility in solution, which hinders the validation of current models. As a consequence, concerns exist in their real generality and physical sense and in whether a small improvement compensates for an increase in model complexity and the need for adjusting more ad hoc parameters.

On the basis of the previous paragraph, it seems reasonable to use MD simulations as reference data for refinement of ENMs. Surprisingly, only a few authors have explored the use of MD data for ENM parametrization. MD simulations render detailed flexibility information on the correlated motions for the time scale sampled, as shown in comparisons with NMR fast motions.³⁶ Current MD simulations reproduce accurately high-quality direct NMR information on protein flexibility (RDCs and S^2 parameters) in the few proteins for which these measurements are available.³⁷ On the other hand, MD displays excellent correlation with B factors, even though MD B factors are systematically larger, especially for very flexible residues (which appear “frozen” in the crystal

lattice³⁷). MD captures both short- and large-scale flexibilities, the latter being extracted from ED treatment of the collected trajectory,³⁸ allowing the characterization of collective anharmonic displacements often related to function.^{39–41} As pointed out above, these so-called *essential* modes also correlate extremely well (both in directions and variance distribution) with the principal components from selected NMR ensembles³⁵ and thus can be expected to provide a quite realistic picture of large-scale flexibility in solution.

In a previous related work,⁴² we performed a thorough comparison between the collective motions predicted by ED and different ENMs. We found that the space defined by the first, most relevant NMA eigenvectors captures the backbone flexibility as given by ED, with the inverse function proposed by Kovacs outperforming the original cutoff approach. However, despite these good correlations with the ED eigenspace, the main motions in NMA are often spread out into multiple modes of similar energy, instead of being concentrated in a few modes as detected in ED. In other words, ED displays higher flexibility, describing collective motions in fewer modes than NMA. Note that this discrepancy cannot be corrected by scaling uniformly the spring constants, since the variance distribution pattern along the energy spectra is fundamentally different (see the discussion below). In this paper we tried to find solutions to this problem by deriving a refined EN-NMA model based on comparison with atomistic MD simulations for a large number of proteins. The proposed ED-refined ENM method (in the following ed-ENM) provides results closest to those of MD, is able to reproduce flexibility in NMR ensembles, and can trace efficiently biologically relevant deformations observed in the Protein Data Bank. The ed-ENM is freely available through the Web site⁵⁸ <http://mmb.pcb.ub.es/Flexserv>. Improvement with respect to standard elastic network models is consistent in all the metrics considered.

2. Methods

2.1. Elastic Network Normal Mode Analysis. ENM can be considered a generalization of the bead-and-strings Rouse polymer model, but contrary to this simple scheme where only chained monomers are coupled, current ENMs connect all α -carbons within a given threshold. Thus, all interactions within the cutoff are harmonic and uniform (irrespective of their chemical nature), and all interactions outside are negligible. By relying on the Cartesian distance as the sole criterion, ENMs are not able to distinguish between close chain neighbors and remote contacts. To derive a more physically sound model, we explored alternative approaches, where the C^α – C^α interaction strength depends on their topological relationship. After extensive testing of different potential functionals, we analyzed in detail three models that represent increasing levels of topological complexity and constant scaling: (i) a cutoff model with a uniform constant, the most widespread approach (standard defaults in ref 43); (ii) a noncutoff model using an exponential decay function, as developed by Kovacs and co-workers;¹⁹ (iii) a hybrid cutoff model with sequential weighted springs for the first

(M) neighbors, while the rest are represented by an inverse function of the Cartesian distance.

To obtain the weights of the spring constants for the first sequential neighbors in an unbiased way, we computed the residue–residue “apparent” stiffness constants obtained from MD assuming the harmonic oscillator model:

$$K_{ij}^{\text{app}} = \frac{k_B T}{\langle [R_{ij} - R_{ij}^0]^2 \rangle} \quad (4)$$

where k_B is the Boltzmann constant, T is the temperature, and $R_{ij} - R_{ij}^0$ is the oscillation in the interaction distance from average values. These constants were fitted to an inverse exponential function using a nonlinear regression routine for a small protein set (see the Results and Discussion):

$$K_{ij}^{\text{app}}(S_{ij}) = \frac{C^{\text{seq}}}{S_{ij}^{n^{\text{seq}}}} \quad (5)$$

where S_{ij} stands for the distance in sequence between residues i and j . The optimum exponent determining the shape of the variation is used in the rest of the study, while the constant C^{seq} is further refined to match “real” instead of “apparent” force constants (see the Results and Discussion). A similar strategy was used to derive the distance dependence for nonsequential interactions:

$$K_{ij}^{\text{app}}(d_{ij}) = \frac{C^{\text{cart}}}{d_{ij}^{n^{\text{cart}}}} \quad (6)$$

where d_{ij} is the distance between residues i and j in a given conformation; in our implementation $d_{ij} = |R_{ij}^0|$. In the ed-ENM, the network topology is defined by a fully connected matrix for the first M neighbors, and contrary to standard pure continuum methods, we introduce a size-dependent cutoff to annihilate artifactual distant interactions (see the Results and Discussion). Thus, given a pair of residues i and j with sequential distance $S_{ij} > 0$ and Cartesian distance d_{ij} , the ij th element of the hybrid inter-residue contact matrix is

$$\Gamma_{ij} \begin{cases} S_{ij} \leq M \\ S_{ij} > M \end{cases} \begin{cases} \Gamma_{ij} = 1 \\ \Gamma_{ij} = 1 \text{ if } d_{ij} \leq r_c \\ \Gamma_{ij} = 0 \text{ otherwise} \end{cases} \quad (7)$$

where Γ_{ij} always has $2M + 1$ nonzero diagonal entries defining neighbor chained contacts. Accordingly, the force constants K_{ij} are dependent not only on the Cartesian but also on the sequential distance:

$$K_{ij} \begin{cases} S_{ij} \leq M \\ S_{ij} > M \end{cases} \begin{cases} K_{ij} = C^{\text{seq}}/S_{ij}^{n^{\text{seq}}} \\ \text{if } d_{ij} \leq r_c \text{ then } K_{ij} = (C^{\text{cart}}/d_{ij})^{n^{\text{cart}}} \\ K_{ij} = 0 \text{ otherwise} \end{cases} \quad (8)$$

where values for all terms ($n^{\text{seq}} = 2$ and $C^{\text{seq}} = 60$ kcal/(mol $\cdot\text{\AA}^2$); $n^{\text{cart}} = 6$ and $C^{\text{cart}} = 6$ kcal/(mol $\cdot\text{\AA}^2$); in energy units) are obtained by fitting to apparent force constants and structural variance profiles. On the basis of MD simulations, a limit of $M = 3$ was used for sequential interactions, and the cutoff radius (r_c) was found to be dependent on the size (see the Results and Discussion).

2.2. Molecular and Essential Dynamics. MD simulations for several proteins (see above) were titrated, neutralized, hydrated, minimized, heated, and equilibrated for at least 0.5 ns. Trajectories were collected for at least 10 ns using three all-atom force fields (AMBER,⁴⁴ CHARMM,⁴⁵ and OPLS/AA⁴⁶). The three trajectories obtained were combined to create a *metatrajectory* which is expected to collect much of the equilibrium dynamics of proteins (control simulations were also performed considering the individual trajectories). The noise arising from irrelevant short-range vibrations was filtered to obtain large-scale motions by ED:⁷ the MD trajectory snapshots were aligned to the original X-ray reference structure (or the average of the NMR ensemble) to compute a common average structure and used to build a covariance matrix whose diagonalization (PCA) yields a set of eigenvectors and eigenvalues representing the essential, large-scale movements (further details in ref 37). To check the ability of the ENMs to capture deformations happening on longer time scales, we extend several calculations to long (0.1 μ s) or very long (0.5–1 μ s) time scales, using in this case only the AMBER force field as discussed below.

2.3. Training Proteins. Initial training of the model was performed by taking six highly representative proteins (PDB 1I6F, 1PHT, 1AGI, 1JLI, 1BSN, and 1SUR) of different sizes (60–200 residues) which were present in our μ MODEL subset of the MODEL database (<http://mmb.pcbub.es/MoDEL>; see ref 37). Parameters were adjusted using as reference the dynamics metatrajectories described above.

2.4. Test Proteins. The model was first tested against the rest of the proteins (from 32 to 400 residues) contained in the μ MODEL set, composed of 32 proteins representing the main metafolds. Larger and multidomain proteins (PDB 3ADK, 1BUD, 1SSX, 1PPO, 1DUA, 1QLJ, and 1PMI) were added, including some extremely large proteins (1SQC (619 residues), 1E5T (710), and 1J0M (747)) and a multimeric complex (1E9S (2545)). To test the performance of ed-ENM on long time scales and avoid any bias introduced by the length of the simulations, we analyzed extended MD trajectories (0.1 μ s) for 2GB1, 1CEI, 1CQY, and 1OPC, up to the microsecond (0.5–1 μ s) for the last two proteins (1CQY, 1OPC) plus 1UBQ and 1KTE. Long trajectories as well as standard trajectories for large proteins were obtained only with AMBER.

2.5. Comparison Metrics. The ENMs' ability to reproduce MD flexibility was tested considering a wide variety of metrics to cover different aspects.

2.5.1. Relative Deformational Amplitude. The size and complexity of the protein deformation space were characterized by (i) the *structural variance*, (ii) the *number of modes needed to explain 90%* of this variance, (iii) the *variance profile* with respect to the number of modes, (iv) the *“reduced variance”* defined as the variance explained by the first five modes, which for most average-sized proteins accounts for 70–80% of the total variance (see Figure S1 in the Supporting Information; similar findings in ref 31), and (v) the *strength of the softer deformation modes*. Note here that the ED eigenvalues obtained by diagonalization of the Cartesian covariance matrix (describing the mode amplitude) appear in distance units, but can be converted into energy

units (kcal/(mol \cdot \AA^2), i.e., mode strength) for comparison with NMA modes by using

$$K_\nu = \frac{k_B T}{\lambda} \quad (9)$$

where ν stands for a given mode, k_B is Boltzmann's constant, T is the temperature, and λ stands for the associated eigenvalue (in square distance, \AA^2).

2.5.2. Deformational Space Overlap. Hess's metric^{47–49} was used to estimate the similarity of NMA and ED deformation spaces:

$$\gamma_{XY} = \frac{1}{m} \sum_{i=1}^m \sum_{j=1}^m (v_i^{\text{ED}} \cdot v_j^{\text{NMA}})^2 \quad (10)$$

where the indexes i and j stand for the orders of the eigenvectors (v , ranked according to their variance contribution) and m stands for the number of eigenvectors in the “important space”, defined as the minimum number needed to explain a certain variance threshold. We considered here two definitions of the important space to guarantee a representative number of eigenvectors in the calculations: (i) eigenvectors needed to explain the 90% variance ($\gamma_{90\%}$) and (ii) the first 50 eigenvectors (γ_{50}). Additionally, the similarity between the first 10 eigenvectors from the ED and normal mode subspace was computed (γ_{10}). However, the similarity index in eq 10 presents two shortcomings: (i) the similarity increases with the important space size and (ii) the index is not sensitive to the eigenvector permutation. To solve the first, we refer to Hess's indexes to background models using Z_{score} :

$$Z_{\text{score}} = \frac{(\gamma_{\text{AB}}(\text{obsd})) - (\gamma_{\text{AB}}(\text{random}))}{\text{std}(\gamma_{\text{AB}}(\text{random}))} \quad (11)$$

where 500 physically meaningful random models were obtained by diagonalization of a covariance matrix derived from discrete molecular dynamics (DMD) simulations performed using a Hamiltonian containing covalent bonds plus a hard sphere potential.¹⁰ To evaluate the impact of permutation, we computed dot products between eigenvector pairs, determining the difference in rank between the ones showing the largest overlap, and used Perez's similarity index, which weighs the similarity of each pair of eigenvectors by their associated Boltzmann factor (see ref 50):

$$\xi_{\text{AB}} = \frac{2 \sum_{i=1}^{i=z} \sum_{j=1}^{j=z} \left[(v_i^{\text{A}} \cdot v_j^{\text{B}}) \frac{\exp\left\{-\frac{(\Delta x)^2}{\lambda_i^{\text{A}}} - \frac{(\Delta x)^2}{\lambda_j^{\text{B}}}\right\}}{\sum_{i=1}^{i=z} \exp\left\{-\frac{(\Delta x)^2}{\lambda_i^{\text{A}}}\right\} \sum_{j=1}^{j=z} \exp\left\{-\frac{(\Delta x)^2}{\lambda_j^{\text{B}}}\right\}} \right]^2}{\sum_{i=1}^{i=z} \left(\frac{\exp\left\{-2\frac{(\Delta x)^2}{\lambda_i^{\text{A}}}\right\}}{\left(\sum_{i=1}^{i=z} \exp\left\{-\frac{(\Delta x)^2}{\lambda_i^{\text{A}}}\right\}\right)^2} \right)^2 + \sum_{j=1}^{j=z} \left(\frac{\exp\left\{-2\frac{(\Delta x)^2}{\lambda_j^{\text{B}}}\right\}}{\left(\sum_{j=1}^{j=z} \exp\left\{-\frac{(\Delta x)^2}{\lambda_j^{\text{B}}}\right\}\right)^2} \right)^2} \quad (12)$$

where the common displacement (Δx) is selected as the minimum value for which the impact outside the important space is negligible. An additional metric that helps in determining the similarity between MD and NMA-based eigenvectors is the “spread” index by Hinsen:⁵¹

$$s_i = \left(\sum_j j^2 \eta_{ij}^2 - \left(\sum_j j \eta_{ij}^2 \right)^2 \right)^{1/2} \quad (13)$$

where $\eta_{ij} = v_i^A \cdot v_j^B$. Note that for two identical sets of modes $\eta_{ij}^2 \neq 0$ only if the $i = j$ spread becomes equal to 0. Higher values indicate the distribution of the eigenvector i on a larger number of eigenvectors j in the B space.

2.5.3. Relative Distribution of the Deformational Pattern. The flexibility distribution along the residues can be analyzed from different metrics. A powerful one is Brüschweiler’s “collectivity” index,⁵² which evaluates the amount of residues involved in every motion k :

$$\kappa_k = \frac{1}{N} \exp \left\{ - \sum_{i=1}^N u_{k,i}^2 \log u_{k,i}^2 \right\} \quad (14)$$

where N is the total number of residues in the protein and $u_{k,i}^2$ is given by

$$u_{k,i}^2 = \frac{v_{k,X}^2 + v_{k,Y}^2 + v_{k,Z}^2}{m_i} \quad (15)$$

where m_i is the mass of residue i . The large-scale motions tend to be the more collective ones. The B factors for each residue i , B_i , were evaluated from average thermal fluctuations, $\langle \Delta r_i^2 \rangle$, under mode k :

$$B_i = (8\pi^2/3) \langle (\Delta r_i)^2 \rangle$$

where

$$\langle (\Delta r_i)^2 \rangle = (3k_B T / \xi) [\Gamma^{-1}]_{ii} = (3k_B T / \xi) \sum_k [\lambda_k^{-1} v_k v_k^T]_{ii} \quad (16)$$

They were also processed to determine Lindemann’s indexes,⁵³ a useful metric providing information on the macroscopic behavior (liquid or solid) of proteins:

$$\Delta_L = \frac{\left(\sum_i \langle \Delta r_i^2 \rangle / N \right)^{1/2}}{a'} \quad (17)$$

where a' is the most probable nonbonded near-neighbor distance (taken as 4.5 Å). To avoid noise introduced by high-frequency modes, B factors and Lindeman’s indexes have been computed by summing the contributions of the first 50 modes (negligible differences are expected if more modes are considered).

2.5.4. Dot Product against X-ray Transition Vectors. Systems selected for analysis belong to a benchmark of conformational transitions (<http://sbg.cib.csic.es/Software/NMAFIT>), formed by 54 transition problems from the macromolecular motions database MolMovDB,⁵⁴ with displacements greater than 2 Å C $^\alpha$ rmsd (the average displacement was 6.3 Å with a standard deviation of 3.4 Å). We

present results for 10 different motions between open/closed pairs; note that each open/closed pair presents two different transition problems. The ability of ed-ENM to predict these biologically relevant transitions was estimated by the accumulated normalized dot products between the 5 (γ_5) and 10 (γ_{10}) first eigenvectors of the corresponding closed/open form, which have been shown to describe the conformational change^{31,54} with respect to the multidimensional vector driving the transition (see eq 10; here $m = 1$ for the first subspace; thus, here γ denotes a dot product between a single vector and a subspace, as opposed to the deformational space overlap in section 2.5.2). As an additional metric, we determined the rank distance between the transition vector and the best overlapped eigenvector (a value of 0 indicates that it is the first one).

2.5.5. Dot Product against Principal Components from NMR Ensembles. To have a qualitative approximation to the flexibility present in NMR multiple structures, we selected 26 ensembles from the Protein Data Bank having at least 10 conformers and spanning a wide size range. Each structure was coarse-grained to the C $^\alpha$ level and then aligned to its average. The closest structure to this initial one was used as a template for a second alignment and computation of the final average structure, which was the reference for subsequent ANM and PCA. The performance of the different ENMs to describe the diversity of the structural ensemble is measured by the accumulated normalized dot products (as given in eq 10) between the 5 (γ_5) and 10 (γ_{10}) first eigenvector pairs from each subspace (i.e., a deformational space overlap as in section 2.5.2) and also by the value of the dot product for the best overlapped pair (γ_{\max}) (a vector to vector inner product).

3. Results and Discussion

3.1. Optimization of the Method. As described above, MD trajectories of a small set of proteins were used to formulate the model, which was later tested against a larger set. The key elements to explore in the training phase were (i) the function for the force constant distance dependence, (ii) the effects of disconnecting/connecting sequential and spatial relationships, (iii) the optimal threshold for distant interactions, and (iv) the pre-exponential factors (C^{scq} and C^{cart} ; see eq 8 for an explanation) used to scale residue–residue stiffness. Our purpose was to define a limit for relevant contacts to infer general connectivity rules. Nevertheless, a multiparametric fitting of all these elements to MD may lead to an overtrained method, and thus, we decided to follow a conservative stepwise strategy to guarantee its generality and physical sense.

3.1.1. Definition of a Sequential Threshold for Nearest-Neighbor Interactions. We first analyzed the distance dependence of the apparent inter-residue force constant detected in MD. The results in Figure 1 (left) show that in the limit of uncoupled oscillators (see eq 4) the apparent force constant decays exponentially with the C $^\alpha$ –C $^\alpha$ distance; similar findings were obtained by Hinsen et al.¹⁸ However, there are evident deviations at distances corresponding to $i \rightarrow i + 1$ residue interactions (close to 3.8 Å) and to a lesser extent

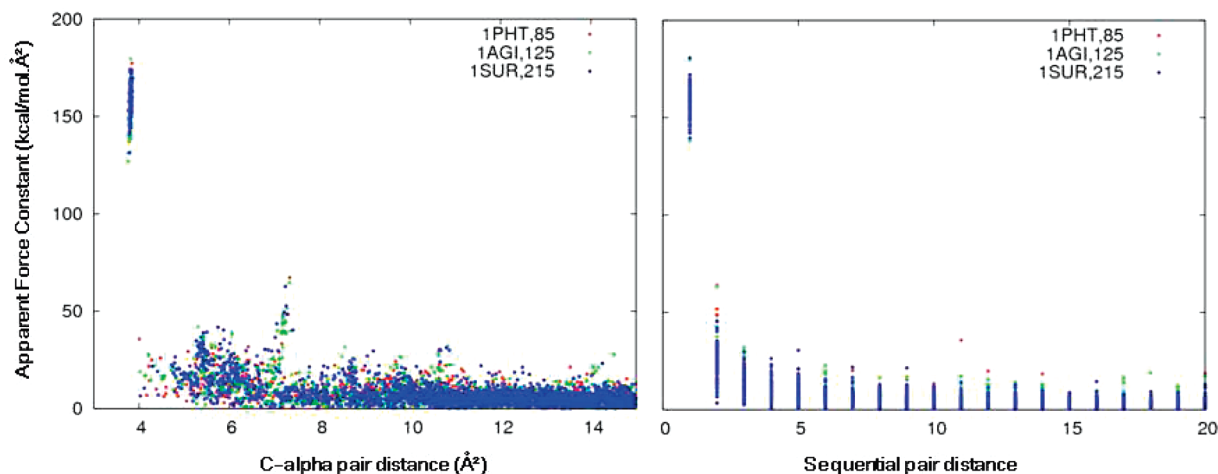


Figure 1. Dependence of the apparent residue–residue force constant (K_{ij}^{app} ; see eq 4) on the residue–residue distance in Cartesian (left, Å) and sequence (right) space as determined for MD in three proteins of different sizes from the training set (see the Methods).

to $i \rightarrow i + 2$ and $i \rightarrow i + 3$ sequence interactions (around 6–10 Å). The singular nature of these nearest-neighbor interactions becomes evident in a plot of the apparent force constant dependence on the sequential distance (Figure 1, right). Fitting of force constants to this sequence distance for close chain neighbors reveals an order 2, inverse square exponential relationship ($n_{\text{seq}} = 2$ in eq 8) which has been incorporated into the algorithm for $i \rightarrow i + 1$ to $i \rightarrow i + 3$ contacts. This formalism defines the relative strength of the interactions between the first three neighbors as approximately $10^2:10^1:10^0$ (in order of magnitude), a ratio that we found important to capture mode directionality. The pre-exponential factor (C^{seq}) appearing in eq 8 cannot be taken directly from MD apparent force constant profiles in Figure 1 and must be refitted to avoid over-restriction of the movement (see the discussion below). To further explore the effects of connectivity, other definitions of the chained residues were analyzed in simpler networks, where an increasing range of sequential contacts was weighted over a background of binary 1/0 contacts for cutoffs from 7 to 25 Å, confirming the $i, i + 3$ limit for main chain interactions (see Figure S3 in the Supporting Information). These simple networks also show that the overlaps follow a peak distribution around maximal values (from 8 to 15 Å) which becomes wider and shifts to higher cutoffs as the chain length increases. It is worth note that the one-neighbor sequence list (topologically equivalent to the constants scaling as 100:1:1 proposed by Jeong et al.²²) gives suboptimal results, suggesting that $i \rightarrow i + 2$ and $i \rightarrow i + 3$ backbone contacts must be clearly weighted over the background defined by a cutoff of ≥ 8 Å. The extension of the sequential singularity to $i \rightarrow i + 5$ interactions did not yield any improvement, as could be anticipated from Figure 1. Interestingly enough, the deletion of distant sequential interactions in a fully connected, continuous network had negligible effects (see Figure S4, top, in the Supporting Information), and conversely, a subminimal NMA model, where only sequential-based $i \rightarrow i + 1$ to $i + 3$ level interactions were included, provided a quite striking agreement with ED modes (Figure S4, bottom). These findings suggest that interactions between sequence neighbors (related to torsional angles

defining the secondary structure) are very important to define the preferred directions of large-scale motions, and therefore, proteins behave as robust networks of reduced connectivity regarding their near-equilibrium dynamics.

3.1.2. Scaling of the Force Constant Energies and the Distance Threshold for Spatial Interactions. When sequential interactions are removed from Figure 1 (right), the apparent force constants are found to decay with the distance following an order 6 exponential ($n_{\text{cart}} = 6$ in eq 8). This sixth-order inverse power law mirrors the distance dependence of the weak, long-range electrostatic interactions determining the 3D fold. Such a dependence, previously proposed by other authors,^{18,19,23} was incorporated into the method, whereas the pre-exponential factor (C^{cart} in eq 8) was further refined against structural variance plots to scale the energy; a size-dependent distance cutoff was introduced to avoid over-restraint of the motions (see the discussion below). In summary, the ed-ENM model treats the strong covalent interaction between nearest-neighbor residues with an order 2 sequentially decaying power law, whereas long-range contacts follow the well-known sixth power law. Once this optimal function was determined, we fitted the force constants by comparison with ED estimates of the (i) total variance, (ii) variance profile, and (iii) reduced variance in order to scale the amplitude distribution of the modes. As mentioned above, we explored values for the sequential, C^{seq} (in the range of 40–200 kcal/(mol·Å²)), and Cartesian, C^{cart} (in the range of 2–12 kcal/(mol·Å²)), constants, finding optimal agreement in the training set for $C^{\text{seq}} = 60$ kcal/(mol·Å²) and $C^{\text{cart}} = 6$ kcal/(mol·Å²). The results are robust to changes of ± 10 kcal/(mol·Å²) in C^{seq} and ± 1 kcal/(mol·Å²) in C^{cart} , particularly regarding the mode directions (see Figure S5 in the Supporting Information).

Analysis of Figure 1 and inspection of the ED of training trajectories reveal that there is a threshold distance from which the apparent restriction in the movement of two pairs of residues is very small and can be explained only by indirect interactions (see Figure S2 in the Supporting Information). This recommends the use of a cutoff to eliminate restrictions to protein movement due to distant negligible interactions. We systematically compared the

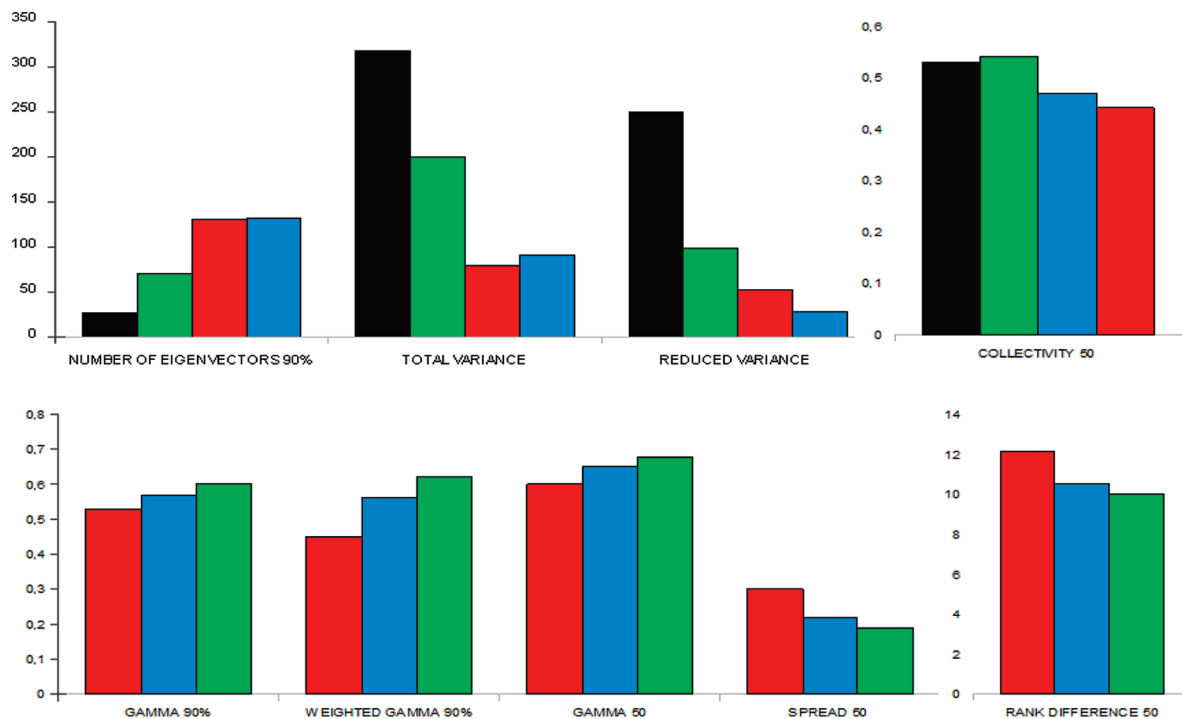


Figure 2. Different metrics for the comparison between MD and ENM-NMA: black, reference MD simulations; green, present ed-ENM model; red, standard cutoff model; blue, Kovacs's formalism.

cutoff, continuous, and mixed cutoff/continuous approaches for a wide range of threshold distances (2–25 Å). Optimum cutoff values were determined by analyzing the dot products between ENMs and ED modes (eq 10) and the relative variance profiles, the magnitudes greatly independent of the value of the force constants. The best results were obtained with mixed continuous/cutoff approaches, and the optimum distance threshold was found to be roughly dependent on the protein size. The size dependence of the cutoff is also clear in the simpler networks tested in Figure S3 in the Supporting Information. We found that the optimal cutoff can be formulated as an approximate logarithmic function of the chain length, starting with a minimal value of 8 Å for the smallest proteins (see the function in Figure S6 in the Supporting Information). This soft size-dependent cutoff (resulting in a practical range from 10 to 16 Å for average to big proteins) removes irrelevant contacts, without affecting important structural details. Subsequently, we will always use this automatic procedure for the cutoff definition in our ed-ENM, avoiding then an arbitrary selection for each protein.

3.2. Validation of the Method. *3.2.1. Validation against MD Flexibility Data for a Representative Benchmark.* As described above, we analyzed the behavior of the new model by comparing our ed-ENM with MD metatrajectories in an extended set of proteins. The reference standard methods used for comparison were the original cutoff approach and the sixth-power exponential function developed by Kovacs et al. (see the Methods). In all cases NMA calculations were performed by taking the MD-averaged structure as a reference to allow direct comparison between the normal modes and MD. All ENMs considered reproduce the ED flexibility pattern reasonably well. Average results for the full μ MODEL set displayed in Figure 2 and more detailed results for

representative proteins in Table 1 illustrate that any of the ENMs are able to capture the overall features of the MD samplings as expected. Similarity indexes with respect to ED (for 90% variance; see eq 10) are in the range of 0.5–0.6 (0.6–0.7 if the index is computed considering 50 eigenvectors), with highly significant associated Z_{score} values (see Table 1 and Figures S7 and S8 in the Supporting Information). These similarity indexes are in fact not far from those obtained by comparing MD trajectories from different force fields among them in the range of 0.7–0.8 (see Table S1 in the Supporting Information; see also ref 54). However, there are considerable differences in the performance of the different methods, and the ed-ENM leads to a moderate but significant increase of around 3–5% in the average similarity index, from 0.54/0.57 to 0.60 in $\gamma_{90\%}$ and from 0.61/0.65 to 0.68 for γ_{50} . Most noticeably, the greatest improvement when using ed-ENM is centered on the prevalent eigenvectors, as shown by the variance-weighted Perez similarity index (see eq 12) for the 90% threshold, which increases from 0.45/0.56 to 0.62 (see $\gamma_{90\%}$ in Table 1), and similar increases in the raw overlap between the eigenspaces defined by the first 10 low-frequency modes (see γ_{10} in Table 1). This close correspondence between MD and ed-ENM lowest frequency motions also becomes clear in the corresponding spread values, lower than those obtained with standard methods (see the average in Figure 2, bottom, and profiles for a few proteins, Figure 4, left).

Analysis of total variances and variance profiles reveals some of the most serious shortcomings of the standard ENMs. First, they underestimate the MD total variance (by a factor of 3–4-fold; see Table 1 and Figure 2, top right), which means that in ENM samplings the structure is too rigid, and this cannot be detected when using crystal flexibility as a reference. Note that the ENM-MD deviation

Table 1. Comparative Measurements of Flexibility Patterns Obtained with NMA and ED of Selected Proteins

PDB code (CATH)	total variance ^a	no. of eigenvectors ^a (90% variance)	similarity (γ_{10}) ^c	similarity ($\gamma_{90\%}$) ^{b,c}	Z _{score} ^c (90% variance)	similarity (γ_{50}) ^c	Z _{score} ^c (50 eigenvectors)	Pearson coefficient ^{c,d}
1OPC 99 (α)	201/67/56/140	19/46/96/44	0.46/0.49/0.48	0.56/0.59/0.61 0.49/0.60/0.60	26/29/31	0.63/0.68/0.70	93/104/109	0.50/0.59/0.65 0.33/0.25/0.39
1CSP 67 (β)	86/45/46/73	20/30/61/38	0.51/0.54/0.61	0.62/0.64/0.68 0.61/0.68/0.72	37/39/44	0.64/0.70/0.72	64/75/79	0.46/0.55/0.71 0.49/0.54/0.62
1SDF 67 ($\alpha + \beta$)	460/76/92/556	7/15/38/9	0.48/0.53/0.53	0.43/0.43/0.49 0.16/0.22/0.52	23/23/28	0.66/0.63/0.67	48/43/50	0.76/0.77/0.79 –
1OOI 124 (α)	131/38/53/103	37/131/133/74	0.28/0.36/0.40	0.59/0.66/0.68 0.47/0.21/0.69	20/34/38	0.63/0.71/0.72	127/149/151	0.40/0.61/0.60 0.23/0.46/0.65
1BFG 126 (β)	85/27/52/75	54/166/143/94	0.44/0.49/0.51	0.62/0.67/0.71 0.66/0.73/0.74	37/50/61	0.62/0.66/0.70	145/158/170	0.39/0.58/0.59 0.30/0.30/0.50
1CHN 126 ($\alpha + \beta$)	359/138/71/160	15/29/118/62	0.46/0.47/0.49	0.48/0.52/0.53 0.38/0.52/0.55	19/23/24	0.61/0.66/0.68	131/146/151	0.54/0.68/0.74 0.35/0.62/0.53
1IL6 166 (α)	840/43/105/252	9/164/139/77	0.50/0.50/0.49	0.49/0.50/0.50 0.09/0.28/0.43	27/28/28	0.60/0.66/0.66	95/109/109	0.68/0.81/0.83 –
1CZT 158 (β)	197/42/112/146	38/140/140/97	0.42/0.49/0.49	0.58/0.65/0.69 0.54/0.70/0.72	42/54/61	0.60/0.65/0.69	111/124/134	0.51/0.56/0.72 0.66/0.67/0.77
1GND 430 ($\alpha + \beta$)	1022/83/248/484	30/521/409/214	0.45/0.51/0.51	0.53/0.56/0.58 0.27/0.32/0.65	23/25/27	0.56/0.61/0.62	330/363/370	0.75/0.77/0.72 0.48/0.57/0.53
1BR5 267 (α)	185/47/150/274	85/353/261/146	0.40/0.44/0.45	0.62/0.68/0.68 0.56/0.73/0.72	41/59/59	0.58/0.64/0.64	200/225/225	0.65/0.71/0.73 –
2PIA 321 (β)	255/69/210/364	96/366/305/162	0.54/0.59/0.60	0.60/0.65/0.66 0.56/0.63/0.71	33/43/46	0.57/0.62/0.62	170/189/189	0.55/0.60/0.62 0.49/0.52/0.50
2HVM 273 ($\alpha + \beta$)	376/32/112/183	44/449/307/184	0.41/0.45/0.45	0.55/0.61/0.60 0.27/0.55/0.61	33/43/42	0.56/0.62/0.61	177/200/196	0.68/0.84/0.81 –

^a Values in the cells always correspond to the MD/cutoff NMA/Kovac/ed-ENM method. ^b Values in the first line of the cells correspond to the standard Hess metrics (eq 10) and values in the second line to the Perez index (eq 12). In every line the results displayed correspond to the cutoff NMA/Kovac/ed-ENM method. ^c Values in the cells correspond to the cutoff NMA/Kovac/ed-ENM method. ^d Values in the first line of the cells correspond to correlations against ED atomic fluctuations and values in the second line to correlations against experimental B factors. In every line the results displayed correspond to the cutoff NMA/Kovac/ed-ENM method.

in variance cannot be fully explained by the fact that we are using an MD metatrajectory as a reference, since it is also evident in single trajectories (see Table S1 in the Supporting Information). Interestingly, the deviation in variance with respect to MD simulations is not uniform for the entire deformation space (which would allow the correction by scaling force constants), but it is larger for the first essential movements, as shown by the reduced variance (see Figure 2, top left). In other words, the MD deformation space is larger (in terms of variance) but less complex (i.e., fewer eigenvectors are required to explain a given variance threshold) than the space described by standard ENMs (see Figure 3, left). The reason for this behavior is clear from the analysis of the variance profiles and the force constants (K_v in eq 9) associated with essential deformations. The standard ENMs and MD simulations distribute variance along the different modes in a different way: while MD defines a small number of soft, highly collective movements which concentrate most of the variance, in ENMs the deformability is distributed along a larger number of eigenvectors. In summary, not only is the total variance different, but MD and standard ENMs also differ in how this variance is partitioned between modes as discussed before, and this is

something that cannot be corrected by scaling a uniform spring constant, since it is more related to the topological properties of the network.

All metrics indicate that ed-ENM yields a remarkable improvement in the total variance and, more important, in the balance of deformation movements as noted in the reduced variance, force constant (K_v in eq 9; Figure 3, right) profiles, and complexity (i.e., number of eigenvectors to capture a certain variance threshold) of the deformation space (see Table 1; μ MoDEL averages in Figure 2, top). It is worth noting that the improvement obtained by using the ed-ENM model is mainly focused on the softest, low-frequency modes and is constant for all the size ranges of proteins considered and for all structural families, as shown by selected examples in Table 1. These soft modes of deformation are highly cooperative, involving a great part of the molecule, as shown by their collectivity degree.⁵² The amount of residues involved in essential movements is similar in ENM and MD according to the Brüschweiler index (0.4–0.6 average for the first 50 modes), but there is a uniform tendency of standard NMA to less collective movements (Figure 2, top right), a situation that is corrected in ed-ENM, possibly due to the strongest nearest-neighbor coupling. Projection of the

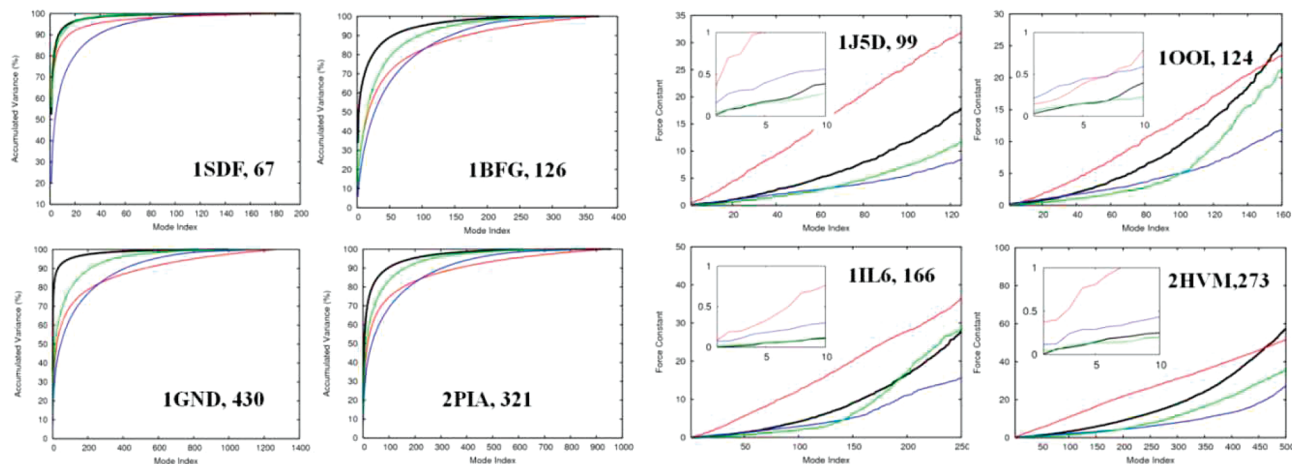


Figure 3. Cumulative variance with respect to the number of eigenvectors (left) and strengths of the essential deformation modes (right, K_i in eq 9) computed by the different methods for some typical proteins (the inset corresponds to a zoom of the first eigenvalues). Illustrative proteins of different sizes and secondary structure compositions are displayed (the name and number of the residues are shown in each graph). The color code is as in Figure 2.

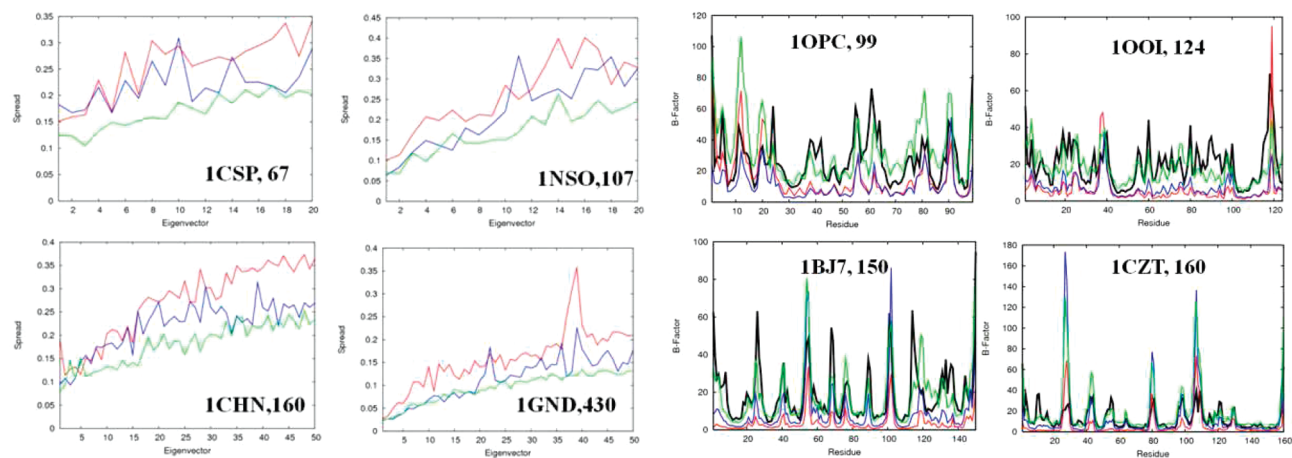


Figure 4. Spread of the eigenvectors in the ED eigenspace for randomly selected proteins (left). B factor profiles (\AA^2) computed by the different methods for a selected number of proteins (right). Values obtained considering in all cases movements along the first 50 eigenvectors. The color code is as in Figure 2.

collective modes on individual residues allowed us to estimate residue fluctuations in solution (see the Methods, eq 16). As previously reported,^{10,42} all ENMs reproduce (see Table 1) the MD atomic fluctuations reasonably well, with Pearson's correlation factors in the range of 0.5–0.6 (Spearman's coefficients typically 0.7–0.8). However, when individual fluctuation distributions are compared (see Figure 4, right), the shortcomings of standard ENMs become evident in a flattening of the profiles, resulting from the problems of ENM in capturing local but large nonharmonic deformations. It is also worth noting that even our interest was not in the description of flexibility in the crystal but that in solution, where the ed-ENM approach also yields a slight improvement (see Figure 4 and Table 1) in the X-ray B factor profiles. We also found that it is possible to raise the correlations for B factors by increasing the distance threshold (unpublished data), but as a result the structure becomes stiffened and the accuracy decreases in other global flexibility measurements, such as the similarity index, variance profiles, or overlap with transition vectors (see below). A simple postprocessing of positional fluctuations allows the derivation of Lindemann's index (see the Methods, eq 17), a key

descriptor to analyze the macroscopic nature of proteins. The results in Table S2 in the Supporting Information illustrate the superiority of the ed-ENM with respect to the standard methods to estimate the absolute MD-derived Lindemann index. The ed-ENM nicely reproduces the core/surface (solid/liquid) asymmetry of proteins and the different macroscopic behavior of the main classes of secondary structures.

3.2.2. Robustness for Large Proteins and Extended Simulations. All results reported to this point suggest that the ed-ENM provides a better approximation of protein flexibility in solution when compared to standard ENM models. There are, however, two reasons for concern that have not yet been addressed regarding the behavior of the model on the biologically relevant time and length scales: (i) What happens when large proteins are considered (larger than those analyzed during the calibration)? (ii) What happens when the new ed-ENM is compared with the flexibility description obtained from long trajectories, where the protein is expected to display larger nonharmonic deformations? To answer the first question, we extended our study to several large and multimeric proteins (from 600 to 2500 residues), finding that ed-ENM captures well their fundamental dynamics (see

Table 2. Rmsd (Å) between X-ray Conformations, Overlaps (%) between Essential Deformation Spaces^a and the Transition Vector, and Rank of Maximum Overlap^b for the Cutoff, the Inverse Exponential Model, and the ed-ENM^c

length (no. of residues) (CATH)	PDB code	rmsd	γ_5^d	γ_{10}^d	rank difference ^{d,e}
101	1L5E (open)	8.8	0.76/0.43/0.81	0.81/0.76/0.85	0 (0.70)/0 (0.66)/0 (0.65)
	1L5B (closed)		0.83/0.80/0.81	0.86/0.85/0.87	1 (0.27)/1 (0.28)/2 (0.55)
148	1CFD (open)	10.2	0.88/0.93/0.94	0.93/0.94/0.95	1 (0.38)/0 (0.62)/1 (0.55)
	1CFC (closed)		0.83/0.89/0.89	0.93/0.92/0.94	1 (0.55)/0 (0.45)/1 (0.55)
214	4AKE (open)	8.3	0.90/0.90/0.92	0.93/0.92/0.93	0 (0.67)/0 (0.38)/0 (0.67)
	1AKE (closed)		0.55/0.57/0.64	0.61/0.68/0.71	0 (0.32)/0 (0.36)/0 (0.40)
219	1NBV (H) (open)	2.2	0.69/0.69/0.70	0.73/0.72/0.73	2 (0.68)/0 (0.29)/0 (0.32)
	1CBV (H) (closed)		0.68/0.69/0.71	0.72/0.71/0.72	2 (0.38)/2 (0.40)/0 (0.37)
271	1URP (open)	7.7	0.96/0.93/0.95	0.96/0.95/0.97	1 (0.94)/1 (0.72)/1 (0.80)
	2DRI (closed)		0.83/0.82/0.88	0.86/0.88/0.92	0 (0.62)/0 (0.56)/1 (0.71)
317	1CKM (A) (open)	4.3	0.93/0.91/0.93	0.94/0.93/0.95	0 (0.86)/0 (0.44)/0 (0.88)
	1CKM (B) (closed)		0.21/0.49/0.57	0.65/0.73/0.78	6 (0.29)/2 (0.14)/4 (0.23)
320	3DAP (open)	5.8	0.89/0.90/0.93	0.94/0.92/0.95	0 (0.75)/1 (0.58)/0 (0.68)
	1DAP (closed)		0.20/0.18/0.27	0.44/0.62/0.78	9 (0.19)/7 (0.33)/4 (0.22)
401	9AAT (open)	2.2	0.15/0.07/0.55	0.68/0.64/0.71	5 (0.26)/5 (0.45)/4 (0.44)
	1AMA (closed)		0.07/0.08/0.60	0.68/0.67/0.76	6 (0.30)/5 (0.39)/6 (0.30)
452	1BNC (open)	5.4	0.84/0.85/0.87	0.87/0.90/0.90	0 (0.83)/0 (0.71)/0 (0.81)
	1DV2 (closed)		0.70/0.69/0.76	0.77/0.80/0.85	4 (0.22)/0 (0.40)/0 (0.48)
517	1RKM (open)	5.8	0.93/0.92/0.93	0.94/0.94/0.95	0 (0.91)/0 (0.84)/0 (0.92)
	2RKM (closed)		0.62/0.64/0.67	0.68/0.75/0.73	1 (0.32)/0 (0.52)/0 (0.42)
avg	open		0.66/0.67/0.70	0.75/0.75/0.76	0.9 (0.69)/0.6 (0.57)/0.6 (0.67)
	closed		0.51/0.53/0.58	0.65/0.67/0.69	3.0 (0.35)/1.7 (0.38)/1.7 (0.42)

^a Considering the first 5 and 10 eigenvectors. ^b See the Methods for a description of the different metrics. ^c Examples from the NMAfit benchmark. ^d Values in these columns are for the cutoff/inverse/ed-ENM. ^e In parentheses, the dot product of the maximal overlap vector is given.

Table S3 and Figures S9 and S10 in the Supporting Information). This confirms that the method can be transferred to analyze large systems, difficult to tackle by MD simulations. The second challenge was to compare the ed-ENM modes to those derived from long MD trajectories (from 0.1 to 0.5–1 μ s), where nonharmonic movements are likely to have more impact on the dynamics. Once again, all the metrics demonstrate the robustness and generality of the ed-ENM, in particular, in correcting the splitting of the soft modes observed in standard approaches (see Table S4 and Figures S11 and 12 in the Supporting Information). However, though the variance descriptors remain at the same order of magnitude, there is a uniform tendency for all ENMs to lower the similarity indexes when the time span of the MD is extended (see similarity index values falling from 0.6–0.7 to 0.4–0.5 for 1CQY and 1OPC in Table S4). This is not surprising since, in a longer trajectory, the structures are able to explore a wider conformational subspace and thus undergo anharmonic departures from equilibrium that cannot be fully captured by any NMA-based approach as discussed above.

3.2.3. Validation against Empirical Flexibility Data from X-ray and NMR. Finally, we tested the method against experimental data on flexibility from both X-ray conformer transitions and PCA of selected NMR ensembles (see the Methods). First, we analyzed the ability of ed-ENM to predict functional important closed/open transitions between X-ray conformers. These large-scale rearrangements involve cooperative motions of domains or subunits, behaving as rigid clusters but preserving the overall fold; in this case the local cohesion prevails over interdomain, long-range interactions. Hence, a great shortcoming in continuum ENM approaches is the over-restriction of displacements between domains, as noticed before.²³ On the other hand, ENM cutoff ap-

proaches display a difficult balance between violation of dihedral constraints for lower distance thresholds and over-restriction of motions if increased. We expected that the combination of a sixth power law with a soft size-dependent cutoff, together with the strongest, inverse-square cohesion limited to neighbors, would allow more natural internal movements. To verify our hypothesis, we studied a benchmark of selected conformational transitions from the macromolecular motion database MolMovDB⁵⁴ (<http://www.molmovdb.org>). Average results for the full benchmark (54 structures) and detailed data for 10 selected cases are displayed in Table 2: 4 structures undergoing large transitions (rmsd > 7 Å) and 6 more with local, less dramatic changes (rmsd = 2–6 Å). The results show that all ENMs encode the functional transitions in their intrinsic flexibility, but the ed-ENM provides the best agreement between the transition vector and the harmonic deformation space. In the open forms, considering only the first 5 modes, the overlaps range from around 0.60 to 0.95 (average 0.7) and from 0.70 to 0.97 (average 0.76) if the harmonic space is extended to 10 modes (see γ_5 and γ_{10} in Table 2); note that random deformations would yield overlaps around 0.08 (5 eigenvectors) and 0.16 (10 eigenvectors). There is a systematic trend to better performance of the ed-ENM (2–5%) regardless of the extent of the transition, particularly remarkable when considering only the first five dominant modes. The greatest improvement using the ed-ENM is achieved for the closed forms, more difficult to treat since they can be easily overconstrained by long-range springs: in this case γ_5 increases by nearly 10% (from 0.50 in standard approaches to almost 0.60). The agreement is particularly surprising in the most challenging cases, where other ENMs fail dramatically (see, for example, the *closed* \rightarrow *open* transition for 1CKM (B), 1AMA, and 1DAP). These notable differences

Table 3. Cumulative Overlaps between the First 10 (γ_{10}) Normal Modes and PCs from NMR Ensembles and Largest Overlap (γ_{\max}) between an ENM Mode and the Best Overlapped PC for Each Set (See Eq 10) for 26 Proteins

PDB code	<i>N</i>	<i>M</i>	$\gamma_{(5)}^a$			$\gamma_{(10)}^a$				γ_{\max}	
1RO4	58	35	0.56	0.52	0.58	0.57	0.53	0.62	0.59	0.33	0.72
1E9T	59	59	0.51	0.48	0.63	0.53	0.54	0.58	0.48	0.57	0.64
1BW5	66	50	0.69	0.59	0.74	0.56	0.56	0.62	0.53	0.63	0.78
2EOT	74	32	0.52	0.42	0.61	0.56	0.41	0.55	0.57	0.76	0.54
1A6X	87	49	0.55	0.44	0.56	0.41	0.37	0.48	0.54	0.37	0.95
1BVE	99	28	0.47	0.52	0.49	0.33	0.36	0.37	0.81	0.7	0.88
1Q06	101	55	0.57	0.58	0.57	0.51	0.60	0.57	0.51	0.58	0.77
2CZN	103	38	0.62	0.55	0.66	0.50	0.53	0.51	0.87	0.65	0.70
1A90	108	31	0.36	0.44	0.49	0.38	0.40	0.42	0.65	0.49	0.83
2BO5	120	44	0.54	0.37	0.58	0.55	0.45	0.56	0.73	0.57	0.68
1E5G	120	50	0.71	0.70	0.69	0.60	0.64	0.63	0.93	0.90	0.96
1CMO	127	43	0.70	0.61	0.70	0.56	0.52	0.59	0.56	0.60	0.52
1ITI	133	31	0.53	0.65	0.59	0.46	0.51	0.44	0.78	0.78	0.89
1C89	134	40	0.70	0.76	0.80	0.55	0.60	0.63	0.53	0.69	0.63
1XSB	153	39	0.46	0.43	0.49	0.44	0.38	0.43	0.89	0.41	0.92
1BF8	205	20	0.47	0.55	0.54	0.43	0.47	0.48	0.86	0.78	0.88
1BY1	209	20	0.55	0.55	0.56	0.42	0.46	0.47	0.50	0.65	0.53
1N6U	212	22	0.63	0.61	0.59	0.58	0.58	0.58	0.64	0.37	0.60
2JZ4	299	20	0.53	0.51	0.61	0.41	0.40	0.45	0.49	0.80	0.74
2D21	370	20	0.60	0.56	0.65	0.50	0.47	0.50	0.67	0.62	0.62
avg			0.56	0.54	0.61	0.49	0.49	0.53	0.65	0.61	0.74

^a Values in these columns are for the cutoff/inverse/ed-ENM.

are related to the concentration of the conformational change in the first dominant eigenvectors. Accordingly, the rank differences are often also smaller and the best overlapped eigenvectors closest to the transition direction. In summary, the ed-ENM displays a higher cooperativity and less dispersion of the motions—as in the above comparison with ED—and thus traces the functional changes with fewer modes.

Finally, we analyzed the ability of ed-ENM to approach the structural diversity of NMR ensembles, which in a first approach can be related (not in a fully rigorous manner) to the experimental flexibility pattern. The analysis of 20 selected NMR multiple structures shows striking correlations with the three ENMs (see Table 3), which confirms previous results³¹ and supports the validity of ENMs to sample the near-equilibrium conformational space in solution. It is also clear that the ed-ENM method outperforms the other two ENM approaches, especially when considering only the first 5 eigenvectors whose overlap γ_5 increases from 0.56/0.54 to 0.61 and the best overlapped pair (γ_{\max}), which increases from a 0.65/0.61 average to 0.74, reaching values near 0.90 (see 1BVE, 1ITI, and 1BF8) or even above (1A6X, 1E5G, and 1XSB). In more than half of the proteins (11 cases), the best overlapped vector is found in the ed-ENM method, followed by the inverse (5 cases) and cutoff (4 cases) approaches, following the trend observed in the rest of the tests. In conclusion, the ed-ENM seems to provide a significant and systematic improvement in the description of protein dynamics (as deduced from structural diversity in NMR ensembles) with respect to the two most used ENM implementations.

4. Conclusions

The ability of the elastic network NMA models to predict qualitatively the intrinsic motions of proteins has been widely demonstrated in the past few years. In comparison with MD,

ENMs tend to yield a sparser pattern of flexibility, related to their harmonic character, and then, the information required for a realistic description of a functional motion is dispersed into a higher number of modes.⁴² Another problem of ENMs has been the lack of consensus in the refinement, mainly due to the scarcity of direct measurements of protein flexibility. In previous studies we demonstrated that MD gives an accurate picture of flexibility in solution.³⁷ In this work we have used atomistic simulations as an alternative source for ENM refinement to extract connectivity rules and obtain a realistic scaling of the force constants. These constraints led to the formulation of a new ED-refined ENM (ed-ENM), based on a simple hybrid potential considering chain topology, which has been validated against a database of MD trajectories. The method proposes a simple and robust scaling of the local backbone and long-range contacts, avoiding any arbitrary, free parameters. A soft size-dependent cutoff is applied to eliminate noise from irrelevant contacts and increase computational efficiency when dealing with large systems. Our goal was not to reproduce any particular flexibility measurement (such as *B* factor profiles), but rather to develop a general method able to trace protein flexibility better than or at least as well as the best performing standard approach for the widest range of descriptors and the largest variety of protein sizes and folds. As discussed above, higher scores for individual flexibility measurements can be achieved by problem-specific adjustment of the ENMs, but only compromising accuracy in other aspects. For example, large cutoffs boost correlations with *B* factors, but the resulting more rigid structures cannot display large conformational transitions. Clearly, when considering all the flexibility measurements presented here, the ed-ENM outperforms standard approaches in the representation of both local and global flexibility for a wide range of proteins and without any ad hoc adjustments. Comparisons to submicrosecond MD suggest that ed-ENM is flexible enough to partially capture

nonharmonic deformations. The method is robust, general, and transferable and can describe large conformational transitions required for biological activity. Finally, we have demonstrated that the method introduced here captures the flexibility of NMR structural ensembles with remarkable precision. Therefore, the bulk of results presented demonstrate that the ED-refined ENM can be a useful alternative to well-established coarse-grained NMA methods. The ability of a minimalist model based on close-chain neighbor interactions both to match molecular dynamics and to trace these complex transitions strongly supports the hypothesis that local covalent topology encodes an important part of the intrinsic flexibility pattern of proteins and thus guides biologically relevant conformational changes. Though this idea may appear somewhat counterintuitive, given the importance of long-range interactions for the 3-D fold, it is just outlining the fact that conformational transitions usually involve motions of rigid residue clusters that maintain the local fold and that this local fold is dependent on the nearest-neighbor contacts, stereochemically restrained. Thus, not only the global contact topology^{11,54–57} but also the inner topology defined from nearest-neighbor contacts plays a great role in the determination of these lowest energy, intrinsically favored modes.

Acknowledgment. This work was supported by the Spanish Ministry of Education and Science (Consolider E-Science and Grant BIO2009-10964), the Spanish Ministry of Health (COMBIOMED project), the Fundación Marcelino Botín, and the National Institute of Bioinformatics. Calculations were performed on the MareNostrum supercomputer at the BSC. L.O. is funded by a predoctoral fellowship from the Spanish Health Ministry.

Supporting Information Available: Additional tables containing comparative metrics obtained with different force fields and the ed-ENM (Table S1), a comparison of Lindemann's coefficients (Table S2), and comparative metrics for large proteins and submicrosecond MD simulations to evaluate robustness in extended length (Table S3) and time (Table S4) scales and additional figures giving the percentage of variance captured by the first five modes in the benchmark (Figure S1), an illustration of the difference between direct and indirect interactions (Figure S2), similarity index profiles in sequential-based networks (Figure S3), similarity index profiles switching on/off the sequential constants (Figure S4), the robustness to changes in Cartesian/sequential constants (Figure S5), the size-dependent cutoff function (Figure S6), comparative similarity index and Z_{score} values for the benchmark (Figures S7 and S8), variance profiles and force constants for large proteins (Figures S9 and S10), and extended submicrosecond MD simulations (Figures S11 and S12). This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- (1) Henzler-Wildman, K. A.; Lei, M.; Thai, V.; Kerns, S. J.; Karplus, M.; Kern, D. A hierarchy of timescales in protein dynamics is linked to enzyme catalysis. *Nature* **2007**, *450*, 913–916.
- (2) Velázquez-Muriel, J. A.; Rueda, M.; Cuesta, I.; Pascual-Montano, A.; Orozco, M.; Carazo, J. M. Comparison of molecular dynamics and superfamily spaces of protein domain deformation. *BMC Struct. Biol.* **2009**, *9*, 6.
- (3) Lindorff-Larsen, K.; Best, R. B.; Depristo, M. A.; Dobson, C. M.; Vendruscolo, M. Simultaneous determination of protein structure and dynamics. *Nature* **2005**, *433*, 128–132.
- (4) McCammon, J. A.; Gelin, B. R.; Karplus, M. Dynamics of folded proteins. *Nature* **1977**, *267*, 585–590.
- (5) Dynamical simulation methods. In *Proteins: A Theoretical Perspective of Dynamics, Structure and Thermodynamics*; Brooks, C. L., III, Karplus, M., Pettitt, B. M., Eds.; Advances in Chemical Physics, Vol. LXXI; John Wiley & Sons Ltd.: New York, 1988; pp 33–58.
- (6) Karplus, M.; Kuriyan, J. Molecular dynamics and protein function. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6679–6685.
- (7) Amadei, A.; Linssen, A. B.; Berendsen, H. J. Essential dynamics of proteins. *Proteins: Struct., Funct., Genet.* **1993**, *17*, 412–425.
- (8) Tozzini, V. Coarse-grained models for proteins. *Curr. Opin. Struct. Biol.* **2005**, *15*, 144–150.
- (9) Bahar, I.; Rader, A. J. Coarse-grained normal mode analysis in structural biology. *Curr. Opin. Struct. Biol.* **2005**, *15*, 586–592.
- (10) Emperador, A.; Carrillo, O.; Rueda, M.; Orozco, M. Exploring the suitability of coarse-grained techniques for the representation of protein dynamics. *Biophys. J.* **2008**, *95*, 2127–2138.
- (11) Tirion, M. M. Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. *Phys. Rev. Lett.* **1996**, *77*, 1905–1908.
- (12) Case, D. A. Normal mode analysis of biomolecular dynamics. In *Computer Simulation of Biomolecular Systems*; Gunsteren, W. F., Weiner, P. K., Wilkinson, A. J., Eds.; Kluwer Academic Publishers: Dordrecht, The Netherlands, 1997; Vol. 3, pp 284–301.
- (13) Atilgan, A. R.; Durell, S. R.; Jernigan, R. L.; Demirel, M. C.; Keskin, O.; Bahar, I. Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophys. J.* **2001**, *80*, 505–515.
- (14) Tama, F.; Gadea, F. X.; Marques, O.; Sanejouand, Y. H. Building-block approach for determining low-frequency normal modes of macromolecules. *Proteins: Struct., Funct., Genet.* **2000**, *41*, 1–7.
- (15) Kondrashov, D. A.; Cui, Q.; Philips, G. N. Optimization and evaluation of a coarse-grained model of protein motion using X-ray crystal data. *Biophys. J.* **2006**, *91*, 2760–2767.
- (16) Chennubhotla, C.; Bahar, I. Markov methods for hierarchical coarse-graining of large protein dynamics. *J. Comput. Biol.* **2007**, *14*, 765–76.
- (17) Sen, T. Z.; Jernigan, R. L. Optimizing the parameters of the Gaussian network model for ATP-binding proteins. In *Normal Mode Analysis: Theory and Applications to Biological and Chemical Systems*; Cui, Q., Bahar, I., Eds.; CRC Press: Boca Raton, FL, 2006.
- (18) Hinsen, K.; Petrescu, A.; Dellerue, S.; Bellissent-Funel, M.; Kneller, G. Harmonicity in slow protein dynamics. *Chem. Phys.* **2000**, *261*, 25–37.

- (19) Kovacs, J. A.; Chacon, P.; Abagyan, R. Predictions of protein flexibility: First-order measurements. *Proteins: Struct., Funct., Bioinf.* **2004**, *56*, 661–668.
- (20) Riccardi, D.; Cui, Q.; Phillips, G. N. Application of elastic network models to proteins in the crystalline state. *Biophys. J.* **2009**, *96*, 464–475.
- (21) Moritsugu, K.; Smith, J. C. Coarse-grained biomolecular simulation with REACH, realistic extension algorithm via covariance Hessian. *Biophys. J.* **2007**, *93*, 3460–3469.
- (22) Jeong, J. I.; Jang, Y.; Kim, M. K. A connection rule for a-carbon coarse-grained elastic network models using chemical bond information. *J. Mol. Graphics Modell.* **2006**, *24*, 296–306.
- (23) Yang, L.; Song, G.; Jernigan, R. L. Protein elastic network models and the ranges of cooperativity. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 12347–52.
- (24) Wagner, G. NMR relaxation and protein mobility. *Curr. Opin. Struct. Biol.* **1993**, *3*, 748–754.
- (25) Gabel, F.; Bicout, D.; Lehnert, U.; Tehei, M.; Weik, M.; Zaccai, G. Protein dynamics studied by neutron scattering. *Q. Rev. Biophys.* **2002**, *35*, 327–367.
- (26) Erman, B. The Gaussian network model: Precise prediction of residue fluctuations and application to binding problems. *Biophys. J.* **2006**, *91*, 3589–3599.
- (27) Hinsen, K. Structural flexibility in proteins: Impact of the crystal environment. *Bioinformatics* **2008**, *24*, 521–528.
- (28) Halle, B. Flexibility and packing in proteins. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 1274–1279.
- (29) Soheilifard, R.; Makarov, D. E.; Rodin, G. J. Critical evaluation of simple network models of protein dynamics and their comparison with crystallographic B-factors. *Phys. Biol.* **2008**, *5*, 026008–026021.
- (30) Carugo, O.; Argos, P. Reliability of atomic displacement parameters in protein crystal structures. *Acta Crystallogr., D: Biol. Crystallogr.* **1999**, *55*, 473–478.
- (31) Tama, F.; Sanejouand, Y. H. Conformational change of proteins arising from normal mode calculations. *Protein Eng.* **2001**, *14*, 1–6.
- (32) Yang, L.; Song, G.; Carriquiry, A.; Jernigan, R. L. Close correspondence between the motions from principal component analysis of multiple HIV-1 protease structures and elastic network modes. *Structure* **2008**, *16*, 321–330.
- (33) Yang, L.; Eyal, E.; Bahar, I.; Kitao, A. Principal component analysis of native ensembles of biomolecular structures (PCA_NEST): Insights into functional dynamics. *Bioinformatics* **2009**, *25*, 606–614.
- (34) Yang, L.; Eyal, E.; Chennubhotla, C.; Jee, J. G.; Gronenborn, A.; Bahar, I. Insights into equilibrium dynamics of proteins from comparison of NMR and X-ray data with computational predictions. *Structure* **2007**, *15*, 741–749.
- (35) Abseher, R.; Horstink, L.; Hilbers, C. W.; Nilges, M. Essential spaces defined by NMR structure ensembles and molecular dynamics simulation show significant overlap. *Proteins: Struct., Funct., Genet.* **1999**, *31*, 370–382.
- (36) Case, D. A. Molecular dynamics and NMR spin relaxation in proteins. *Acc. Chem. Res.* **2002**, *35*, 325–331.
- (37) Rueda, M.; Ferrer-Costa, C.; Meyer, T.; Pérez, A.; Camps, J.; Hospital, A.; Gelpí, J. L.; Orozco, M. A consensus view of protein dynamics. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 796–801.
- (38) Berendsen, H. J. C. Bio-molecular dynamics comes of age. *Science* **1996**, *271*, 954–955.
- (39) Ichiye, T.; Karplus, M. Collective motions in proteins: A covariance analysis of atomic fluctuations in molecular dynamics and normal mode simulations. *Proteins: Struct., Funct., Genet.* **1991**, *11*, 205–217.
- (40) Kitao, A.; Go, N. Investigating protein dynamics in collective coordinate space. *Curr. Opin. Struct. Biol.* **1999**, *9*, 164–9.
- (41) Hayward, S.; Kitao, A.; Go, N. Harmonicity and anharmonicity in protein dynamics: A normal modes and principal component analysis. *Proteins: Struct., Funct., Genet.* **1995**, *23*, 177–186.
- (42) Rueda, M.; Chacón, P.; Orozco, M. Thorough validation of protein normal mode analysis: A comparative study with essential dynamics. *Structure* **2007**, *15*, 565–575.
- (43) Suhre, K.; Sanejouand, Y. H. ElNemo: A normal mode Web server for protein movement analysis and the generation of templates for molecular replacement. *Nucleic Acids Res.* **2004**, *32*, W610–W614.
- (44) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.* **1995**, *117*, 5179–5197.
- (45) MacKerell, A. D.; Wiorkiewicz-Kuczera, J.; Karplus, M. An all-atom empirical energy function for the simulation of nucleic acids. *J. Am. Chem. Soc.* **1995**, *117*, 11946–11975.
- (46) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc.* **1996**, *118*, 11225–11236.
- (47) Hess, B. Similarities between principal components of protein dynamics and random diffusion. *Phys. Rev. E* **2000**, *62*, 8438–8448.
- (48) Noy, A.; Meyer, T.; Ferrer, C.; Valencia, A.; Pérez, A.; de la Cruz, X.; López-Bes, J. M.; Pouplana, R.; Fernandez-Recio, J.; Luque, F. J.; Orozco, M. Data mining of molecular dynamics trajectories of nucleic acids. *J. Biomol. Struct. Dyn.* **2006**, *23*, 447–456.
- (49) Orozco, M.; Perez, A.; Noy, A.; Luque, F. J. Theoretical methods for the simulation of nucleic acids. *Chem. Soc. Rev.* **2003**, *32*, 350–364.
- (50) Pérez, A.; Blas, J. R.; Rueda, M.; López-Bes, J. M.; de la Cruz, X.; Orozco, M. Exploring the essential dynamics of B-DNA. *J. Chem. Theory Comput.* **2005**, *1*, 790–800.
- (51) Hinsen, K. Analysis of domain motions by approximate normal mode calculations. *Proteins: Struct., Funct., Genet.* **1998**, *33*, 417–429.
- (52) Brüschweiler, R. Collective protein dynamics and nuclear spin relaxation. *J. Chem. Phys.* **1995**, *102*, 3396–3403.
- (53) Zhou, Y.; Vitkup, D.; Karplus, M. Native proteins are surface-molten solids: Application of the Lindemann criterion for the solid versus liquid state. *J. Mol. Biol.* **1999**, *285*, 1371–1375.
- (54) Krebs, W. G.; Alexandrov, V.; Wilson, C. A.; Echols, L.; Yu, H.; Gerstein, M. Normal mode analysis of macromolecular motions in a database framework: Developing mode concentration as a useful classifying statistic. *Proteins: Struct., Funct., Genet.* **2002**, *48*, 682–695.

- (55) Emperador, A.; Meyer, T.; Orozco, M. United-atom discrete molecular dynamics of proteins using physics-based potentials. *J. Chem. Theory Comput.* **2008**, *4*, 2001–2010.
- (56) Nicolay, S.; Sanejouand, Y. H. Functional modes of proteins are among the most robust ones. *Phys. Rev. Lett.* **2006**, *96*, 078104.
- (57) Zheng, W.; Brooks, B. R.; Thirumalai, D. Low-frequency normal modes that describe allosteric transitions in biological nanomachines are robust to sequence variations. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 7664–7669.
- (58) Camps, J.; Emperador, A.; Carrillo, O.; Orellana, L.; Hospital, A.; Rueda, M.; Cicin-Sain, D.; D'Abramo, M.; Gelpi, J. L.; Orozco, M. FlexServ: An integrated tool for the analysis of protein flexibility. *Bioinformatics* **2009**, *25*, 1709–10.

CT100208E

JCTC Journal of Chemical Theory and Computation

A Displaced-Solvent Functional Analysis of Model Hydrophobic Enclosures

Robert Abel,[‡] Lingle Wang,[†] Richard A. Friesner,[†] and B. J. Berne^{*†}

*Department of Chemistry, Columbia University, New York, New York 10027,
Schrodinger, L.L.C., New York, New York*

Received April 22, 2010

Abstract: Calculation of protein–ligand binding affinities continues to be a hotbed of research. Although many techniques for computing protein–ligand binding affinities have been introduced—ranging from computationally very expensive approaches, such as free energy perturbation (FEP) theory, to more approximate techniques, such as empirically derived scoring functions, which, although computationally efficient, lack a clear theoretical basis—there remains a pressing need for more robust approaches. A recently introduced technique, the displaced-solvent functional (DSF) method, was developed to bridge the gap between the high accuracy of FEP and the computational efficiency of empirically derived scoring functions. In order to develop a set of reference data to test the DSF theory for calculating absolute protein–ligand binding affinities, we have pursued FEP theory calculations of the binding free energies of a methane ligand with 13 different model hydrophobic enclosures of varying hydrophobicity. The binding free energies of the methane ligand with the various hydrophobic enclosures were then recomputed by DSF theory and compared with the FEP reference data. We find that the DSF theory, which relies on no empirically tuned parameters, shows excellent quantitative agreement with the FEP. We also explored the ability of buried solvent accessible surface area and buried molecular surface area models to describe the relevant physics and find the buried molecular surface area model to offer superior performance over this data set.

I. Introduction

Calculation of relative and absolute protein–ligand binding affinities continues to be an active hotbed of research in the field of computational biophysics.^{1–4} Although many techniques for computing protein–ligand binding affinities have been introduced—ranging from computationally very expensive ab initio approaches, such as free energy perturbation (FEP) theory, to more approximate techniques, such as empirically derived scoring functions, which, although computationally efficient, lack a clear theoretical basis—there remains a pressing need for more robust approaches. A recently introduced technique, the displaced-solvent functional (DSF) method,⁵ was developed to bridge the gap

between the high accuracy of FEP and the computational efficiency of empirically derived scoring functions. This technique proceeds by first using explicitly solvated molecular dynamics simulations of a protein conformation which is complementary to a given ligand series (or, in some cases, a protein–ligand complex which can be used to build the remaining members of the series) to map out the approximate thermodynamic properties of water molecules solvating various regions of the protein active site. Second, a DSF was constructed to compactly represent this information, and third, the relative binding affinities of congeneric ligands were computed for the given receptor by correlating the relative binding affinities of the congeneric ligands with the excess chemical potential of the solvent that is evacuated from the active site by the binding of the ligand.

This method has shown great promise in a number of pharmaceutically relevant applications such as accurately describing the relative binding thermodynamics of proteases,

* To whom correspondence should be addressed. E-mail: bb8@columbia.edu.

[†] Columbia University.

[‡] Schrodinger, L.L.C.

kinases, the PDZ domain, and GPCR inhibitors; elucidating the role of hydration in kinase binding specificity; and offering novel qualitative insights into PCSK9-peptide binding kinetics.^{5–12} However, despite the wide range of successful applications of the technique to describe and explain experimental binding data, the physical–chemical basis of the DSF method has not yet been fully clarified in print. This work derives the DSF approach from first principles and clarifies the physical–chemical basis of the technique. Further, this derivation elucidates the key approximations of the method, which facilitates an understanding of when the technique is expected to succeed and fail. In order to develop a set of reference data to test the DSF theory for calculating absolute protein–ligand binding affinities, we have pursued FEP theory calculations of the binding free energies of a methane ligand with 13 different types of model hydrophobic enclosures of varying hydrophobicities. The binding free energies of the methane ligand with the various hydrophobic enclosures were then recomputed by the DSF theory presented herein, and the results of the calculations were compared with the FEP reference data. We find that the DSF theory predictions, which rely on no empirically tuned parameters, show excellent quantitative agreement with the FEP results (root-mean-square error of 0.40 kcal/mol and an R^2 value of 0.95). Thus, DSF theory may offer, for systems that satisfy the necessary approximations, a method of calculating absolute binding affinities with FEP-like accuracy at only a small fraction of the computational expense. A further point is that the DSF approach can be unambiguously converged with current hardware capabilities, whereas convergence becomes quite challenging for FEP and related methods when applied to complex problems like protein–ligand binding (as opposed to the model systems studied in this paper).

II. Methods

A. Derivation of the Displaced Solvent Functional Approach to Computing Protein Ligand Binding Free Energies. It is well-known¹ that the binding free energy of a small molecule for its cognate protein receptor can be computed as

$$\Delta G_{\text{bind}}^{\circ} = -RT \ln \left[\left(C_0 \int \exp(-[(U(\vec{r}_{\text{PL}}) + W(\vec{r}_{\text{PL}}))/RT]) d\vec{r}_{\text{PL}} \right) / \left(8\pi^2 \int \exp(-[(U(\vec{r}_{\text{P}}) + W(\vec{r}_{\text{P}}))/RT]) d\vec{r}_{\text{P}} \times \int \exp(-[(U(\vec{r}_{\text{L}}) + W(\vec{r}_{\text{L}}))/RT]) d\vec{r}_{\text{L}} \right) \right] \quad (1)$$

where the subscript P represents the protein in the unbound state, the subscript L represents the ligand in the unbound state, the subscript PL represents the protein and ligand in their bound state, R is the gas constant, C_0 is the standard concentration, U is the interaction energy term, and W represents the solvation free energy terms. From this expression one can readily derive

$$\Delta G_{\text{bind}}^{\circ} = \langle U_{\text{PL}} \rangle_{\text{PL}} - \langle U_{\text{P}} \rangle_{\text{P}} - \langle U_{\text{L}} \rangle_{\text{L}} + \langle W_{\text{PL}} \rangle_{\text{PL}} - \langle W_{\text{P}} \rangle_{\text{P}} - \langle W_{\text{L}} \rangle_{\text{L}} - T\Delta S_{\text{config}}^{\circ} \quad (2)$$

where the brackets ($\langle \rangle$) imply Boltzmann weighted averages over the specified ensemble, the changes of the configurational entropies of the protein and the ligand after binding have been grouped in a single term ($-T\Delta S_{\text{config}}^{\circ}$), and the terms related to the change in the interaction energies (U) and solvation free energies (W) of the protein and the ligand are enumerated explicitly. We note here that the $-T\Delta S_{\text{config}}^{\circ}$ term may be made arbitrarily small in eq 2 by first computing the free energy of restraining internal and relative degrees of freedom of the protein and the ligand to some appropriately chosen reference state by FEP, thermodynamics integration, or any other suitable ab initio approach, and then computing the binding free energy of the protein and ligand after these restraints have been applied.^{13,14}

Equation 2, although complete, has poor convergence properties since it is a series of very large terms that sum to a very small number. Thus, each individual term must be computed to very high accuracy and precision. This may in practice be more difficult than sampling eq 1 directly, for example by FEP. However, we have made a series of observations in our recent work^{5,6} that suggest a path to improve the convergence of this expression.

The first observation is that the protein–ligand interaction energy (U_{PL}) can be expanded into an intraprotein term, a protein–ligand interaction term, and an intraligand term:

$$\langle U_{\text{PL}} \rangle_{\text{PL}} = \langle U_{\text{P}} \rangle_{\text{PL}} + \langle U_{\text{P-L}} \rangle_{\text{PL}} + \langle U_{\text{L}} \rangle_{\text{PL}} \quad (3)$$

where the first term (U_{P}) is the intraprotein interaction energy, the second term ($U_{\text{P-L}}$) is the protein–ligand interaction energy, and the third term (U_{L}) is the intraligand interaction energy. Therefore,

$$\Delta G_{\text{bind}}^{\circ} = \langle U_{\text{P}} \rangle_{\text{PL}} + \langle U_{\text{P-L}} \rangle_{\text{PL}} + \langle U_{\text{L}} \rangle_{\text{PL}} - \langle U_{\text{P}} \rangle_{\text{P}} - \langle U_{\text{L}} \rangle_{\text{L}} + \langle W_{\text{PL}} \rangle_{\text{PL}} - \langle W_{\text{P}} \rangle_{\text{P}} - \langle W_{\text{L}} \rangle_{\text{L}} - T\Delta S_{\text{config}}^{\circ} \quad (4)$$

We will assume in this work that the loss of conformational entropy of the protein and ligand is compensated by the strain energy incurred by the protein and ligand upon binding. For example, a ligand with freely rotatable bonds binding to a protein will generally induce little protein strain energy but will lose a great deal of conformational entropy upon binding. Conversely, a highly rigid ligand, which will avoid such entropic penalties, will likely require substantial “induced fit” of the protein, which will in turn increase the strain energy of the protein upon binding. Posed formally, this argument suggests

$$0 \approx \langle U_{\text{P}} \rangle_{\text{PL}} + \langle U_{\text{L}} \rangle_{\text{PL}} - \langle U_{\text{P}} \rangle_{\text{P}} - \langle U_{\text{L}} \rangle_{\text{L}} - T\Delta S_{\text{config}}^{\circ} \quad (5)$$

In turn, eq 4 may be rewritten as

$$\Delta G_{\text{bind}}^{\circ} \approx \langle U_{\text{P-L}} \rangle_{\text{PL}} + \langle W_{\text{PL}} \rangle_{\text{PL}} - \langle W_{\text{P}} \rangle_{\text{P}} - \langle W_{\text{L}} \rangle_{\text{L}} + \delta_{\text{strn}} [\langle U_{\text{P}} \rangle_{\text{PL}} + \langle U_{\text{L}} \rangle_{\text{PL}} - \langle U_{\text{P}} \rangle_{\text{P}} - \langle U_{\text{L}} \rangle_{\text{L}} - T\Delta S_{\text{config}}^{\circ}] \quad (6)$$

where switching function δ_{strn} allows eq 6 to be exact for $\delta_{\text{strn}} = 1$, and approximately correct for $\delta_{\text{strn}} = 0$. Equation 6 may be recognized as equivalent to the MM-GBSA method, where the protein and ligand strain energies and the change in the configurational entropy are neglected when $\delta_{\text{strn}} = 0$, although various formulations have emerged in

the literature.^{15–17} Note, the $\delta_{\text{stm}} = 0$ approximation will be exactly satisfied by the model enclosure studied herein but is expected to apply generally to any series of congeneric ligands binding to a given protein receptor. The reason we expect the $\delta_{\text{stm}} = 0$ approximation to be a reasonable approach to treating a series of congeneric ligands is that small modification of the ligand scaffold can be loosely understood to either make the scaffold slightly more or slightly less rigid, thereby changing the associated entropic cost of the protein binding the ligand. Those modifications that make the ligand more rigid will lead to a less unfavorable binding entropy but will also likely increase the protein strain energy, since the protein must now deform to accommodate a more rigid object. Conversely, small modifications which increase the flexibility of the ligand will reduce the protein strain energy, since less deformation of the protein active site will be required upon binding the ligand but will increase the entropic penalty of the binding process. It is this hypothesized general compensation of the strain energy with the loss of conformational entropy that should lead to the general applicability of the $\delta_{\text{stm}} = 0$ approximate form of eq 6 to congeneric series.

The next series of approximations requires us to restrict our investigations to *complementary ligands*—i.e., ligands that form hydrogen bonds with the protein receptor where appropriate, hydrophobic contacts otherwise, and sterically “fit” within the accessible volume of the active site of the receptor. Such ligands will form interactions with the surrounding protein similar to the interactions the ligand made with the bulk solvent—i.e., hydrogen bonds where appropriate and van der Waals contacts otherwise, be they with the protein active site or with the solvating water. With this in mind, we may rewrite the solvation free energy terms as

$$\langle W_{\text{PL}} \rangle_{\text{PL}} - \langle W_{\text{P}} \rangle_{\text{P}} - \langle W_{\text{L}} \rangle_{\text{L}} = \Delta \langle W_{\text{PL}} \rangle_{\text{P,L;PL}} = \Delta \langle W_{\text{PL}} \rangle_{\text{P,L;PL}}^{\text{cav}} + \Delta \langle W_{\text{PL}} \rangle_{\text{P,L;PL}}^{\text{chrg}} \quad (7)$$

where $\langle W_{\text{PL}} \rangle_{\text{P,L;PL}}$ is the difference in the solvation free energy of the free ligand and protein versus the complex, $\Delta \langle W_{\text{PL}} \rangle_{\text{P,L;PL}}^{\text{cav}}$ is the free energy of growing the repulsive core of the ligand in the bulk versus within the protein active site, and $\Delta \langle W_{\text{PL}} \rangle_{\text{P,L;PL}}^{\text{chrg}}$ is the difference in the free energy of charging the ligand-solvent dispersion and electrostatic interactions in the bulk versus within the protein active site. Such a separation of the charging and cavitation terms is common in FEP studies of protein–ligand binding.^{18,19}

With the introduction of this notation, we find

$$\Delta G_{\text{bind}}^{\circ} \approx \langle U_{\text{P-L}} \rangle_{\text{PL}} + \Delta \langle W_{\text{PL}} \rangle_{\text{P,L;PL}}^{\text{chrg}} + \Delta \langle W_{\text{PL}} \rangle_{\text{P,L;PL}}^{\text{cav}} + \delta_{\text{stm}} [\langle U_{\text{P}} \rangle_{\text{P}} + \langle U_{\text{L}} \rangle_{\text{PL}} - \langle U_{\text{P}} \rangle_{\text{P}} - \langle U_{\text{L}} \rangle_{\text{L}} - T \Delta S_{\text{config}}^{\circ}] \quad (8)$$

We now introduce a rather aggressive approximation

$$\langle U_{\text{P-L}} \rangle_{\text{PL}} \approx -\Delta \langle W_{\text{PL}} \rangle_{\text{P,L;PL}}^{\text{chrg}} + \delta_{\text{sic}} [\langle U_{\text{P-L}} \rangle_{\text{PL}} + \Delta \langle W_{\text{PL}} \rangle_{\text{P,L;PL}}^{\text{chrg}}] \quad (9)$$

where an exact result is obtained for $\delta_{\text{sic}} = 1$, but an approximate result is generated for $\delta_{\text{sic}} = 0$. The rationale for this approximation can be explained as follows:

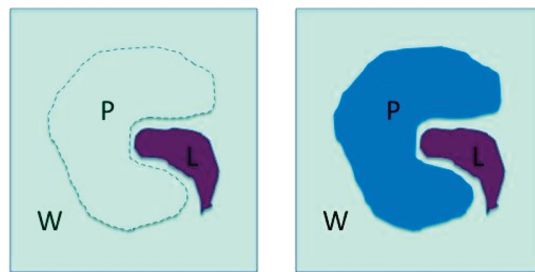


Figure 1. Cartoon depicting the relationship between $\Delta \langle W_{\text{PL}} \rangle_{\text{P,L;PL}}^{\text{chrg}}$ and $\langle U_{\text{P-L}} \rangle_{\text{PL}}$. $\Delta \langle W_{\text{PL}} \rangle_{\text{P,L;PL}}^{\text{chrg}}$ is the free energy difference in turning on the attractive and electrostatic interaction between the ligand and the solvent in the bulk water (left) versus in the active site of protein (right), which is the interaction between the ligand and the solvent that would be excluded by the protein (depicted by dashed line on the left). $\langle U_{\text{P-L}} \rangle_{\text{PL}}$ is the interaction energy between the ligand and the protein in the complex (right). For complementary ligands binding to the protein receptor, the two terms would be expected to be of similar magnitude.

$\Delta \langle W_{\text{PL}} \rangle_{\text{P,L;PL}}^{\text{chrg}}$ is the free energy difference in turning on the dispersion and electrostatic interactions between the ligand and the solvent in bulk water versus in the active site of protein (see Figure 1), which is the interactions between the ligand and the solvent that would be excluded by the protein (depicted by a dashed line in Figure 1); $\langle U_{\text{P-L}} \rangle_{\text{PL}}$ is the interaction energy between the ligand and the protein in the complex (right). For complementary ligands binding to the protein receptor, the two terms would be expected to be similar in magnitude: (1) for polar ligands that make strong interactions with the protein receptor such as a salt bridge, the interaction of the ligands with water would also be strong; and (2) for apolar ligands that make weak dispersion interactions with the protein, the interactions between the ligands and water would also be weak. The reader may wish to note that the approximation described in eq 9 is “aggressive” in the sense that it would be expected to be generally false for an arbitrary ligand binding to an arbitrary receptor. Thus, by employing the approximation described by eq 9, we would only expect the following treatment to well describe ligands that satisfy the underlying assumptions, i.e., that the ligands form hydrogen bonds where appropriate and hydrophobic contacts otherwise. However, with the above caveat noted, we may approximate the binding free energy as

$$\Delta G_{\text{bind}}^{\circ} \approx \Delta \langle W_{\text{PL}} \rangle_{\text{P,L;PL}}^{\text{cav}} + \delta_{\text{sic}} [\langle U_{\text{P-L}} \rangle_{\text{PL}} + \Delta \langle W_{\text{PL}} \rangle_{\text{P,L;PL}}^{\text{chrg}}] + \delta_{\text{stm}} [\langle U_{\text{P}} \rangle_{\text{P}} + \langle U_{\text{L}} \rangle_{\text{PL}} - \langle U_{\text{P}} \rangle_{\text{P}} - \langle U_{\text{L}} \rangle_{\text{L}} - T \Delta S_{\text{config}}^{\circ}] \quad (10)$$

where our identified approximate equivalence between the relative protein–ligand direct interaction energy and the solvation-charging free energies has been explicitly noted in the grouping of the terms. Equation 10 suggests that the binding free energy may be approximated by computing the relative free energies of forming a cavity isosteric to the ligand in the protein active site, versus forming the same cavity in the bulk fluid.

Our remaining task is to develop a computationally efficient procedure to approximate the $\Delta \langle W_{\text{PL}} \rangle_{\text{P,L;PL}}^{\text{cav}}$ term. This term corresponds to the difference in the free energy of

growing the repulsive ligand cavity within the protein active site versus growing the ligand cavity in the bulk, or equivalently dragging the ligand cavity from the bulk through the volume of the system into the active site of the protein. The $\Delta\langle W_{PL} \rangle_{P,L,PL}^{\text{cav}}$ term may be exactly expanded as

$$\Delta\langle W_{PL} \rangle_{P,L,PL}^{\text{cav}} = (G_{\text{IST}}^{\text{PL,cav}} - G_{\text{IST}}^{\text{P}}) - (G_{\text{IST}}^{\text{L,cav}} - G_{\text{IST}}^{\text{H}_2\text{O}(l)}) = \Delta G_{\text{IST}}^{\text{P,PL,cav}} - \Delta G_{\text{IST}}^{\text{H}_2\text{O}(l),\text{L,cav}} = \Delta\Delta G_{\text{IST}}^{\text{L,cav}} \quad (11)$$

where $G_{\text{IST}}^{\text{X}}$ is the inhomogeneous solvation theory²⁰ (IST) integral over the system designated by superscript X, i.e.

$$\begin{aligned} G_{\text{IST}}^{\text{X}} &= E_{\text{IST}}^{\text{X}} - TS_{\text{IST}}^{\text{X}} \\ E_{\text{IST}}^{\text{X}} &= (E^{\text{K}} + E^{\text{sw}} + E^{\text{ww}})^{\text{X}} \\ &= \frac{3}{2}N_w kT + \rho \int g_{\text{sw}}^{\text{X}}(\vec{r}) \times \\ &\quad u_{\text{sw}}^{\text{X}}(\vec{r}) d\vec{r} + \frac{\rho^2}{2} \int g_{\text{sw}}^{\text{X}}(\vec{r}_1, \vec{r}_2) u_{\text{ww}}^{\text{X}}(\vec{r}_1, \vec{r}_2) d\vec{r}_1 d\vec{r}_2 \\ S_{\text{IST}}^{\text{X}} &= (S^{\text{id}} + S^{(1)} + S^{(2)} \dots)^{\text{X}} \\ &= \left[\frac{5}{2}N_w k - kN_w \ln(\rho\Lambda^3) \right] - k\rho \int g_{\text{sw}}^{\text{X}}(\vec{r}) \ln g_{\text{sw}}^{\text{X}}(\vec{r}) d\vec{r} \\ &\quad - \frac{1}{2}k\rho^2 \int g_{\text{sw}}^{\text{X}}(\vec{r}_1, \vec{r}_2) [\ln \delta g_{\text{sw}}^{\text{X}}(\vec{r}_1, \vec{r}_2) - \delta g_{\text{sw}}^{\text{X}}(\vec{r}_1, \vec{r}_2) + 1] d\vec{r}_1 d\vec{r}_2 \dots \\ \delta g_{\text{sw}}^{\text{X}} &= \frac{g_{\text{sw}}^{\text{X}}(\vec{r}_1, \vec{r}_2)}{g_{\text{sw}}^{\text{X}}(\vec{r}_1) g_{\text{sw}}^{\text{X}}(\vec{r}_2)} \quad (12) \end{aligned}$$

where g_{sw} , g_{ww} , and g_{sww} are the solute–water, water–water, and solute–water–water correlation functions; u_{sw} and u_{ww} are the solute–water and water–water interaction energy terms; \vec{r} represents the solvent degrees of freedom of system X; ρ is the density of the bulk fluid; and k is the Boltzmann constant.

Another simplification can be made by noting that the IST integrals appearing in eq 12 can be decomposed into two contributions: the contribution coming from the integral over the space of the ligand cavity and the contribution coming from the integral over the rest of the space. So the ΔG_{IST} integrals appearing in eq 11 (be they in the bulk fluid or the protein active site) can also be decomposed into the corresponding two contributions: (1) the solvation free energies $\sim N_w$ of the water molecules that were formerly solvating the protein active site and are evacuated into solution by the growth of the ligand cavity ($\Delta G_{\text{IST},\text{Nw solv}}$; which comes from the integral over the ligand cavity part) and (2) the contribution from the solvent located at the L cavity surface ($\Delta G_{\text{IST},\text{surf}}$; which comes from the integral over the rest of the space). This decomposition of the total IST integrals into $\Delta G_{\text{IST},\text{surf}}$ and $\Delta G_{\text{IST},\text{Nw solv}}$ terms may be clarified by inspecting the graphical depiction of the decomposition to be found in Figure 2. It is also worth noting that in this notation $\Delta G_{\text{IST}}^{\text{H}_2\text{O}(l),\text{L,cav}} = \Delta G_{\text{IST},\text{surf}}^{\text{H}_2\text{O}(l),\text{L,cav}}$ exactly, since the water is evacuated from a bulk environment to a bulk environment by the growth of the ligand cavity (i.e., $\Delta G_{\text{IST},\text{Nw solv}}^{\text{H}_2\text{O}(l),\text{L,cav}} = 0$ strictly). Therefore,

$$\begin{aligned} \Delta\Delta G_{\text{IST}}^{\text{L,cav}} &= (\Delta G_{\text{IST},\text{surf}}^{\text{P,PL,cav}} + \Delta G_{\text{IST},\text{Nw solv}}^{\text{P,PL,cav}}) - \Delta G_{\text{IST},\text{surf}}^{\text{H}_2\text{O}(l),\text{L,cav}} \\ &= (\Delta G_{\text{IST},\text{surf}}^{\text{P,PL,cav}} - \Delta G_{\text{IST},\text{surf}}^{\text{H}_2\text{O}(l),\text{L,cav}}) + \Delta G_{\text{IST},\text{Nw solv}}^{\text{P,PL,cav}} \\ &= \Delta\Delta G_{\text{IST},\text{surf}}^{\text{L,cav}} + \Delta G_{\text{IST},\text{Nw solv}}^{\text{P,PL,cav}} \quad (13) \end{aligned}$$

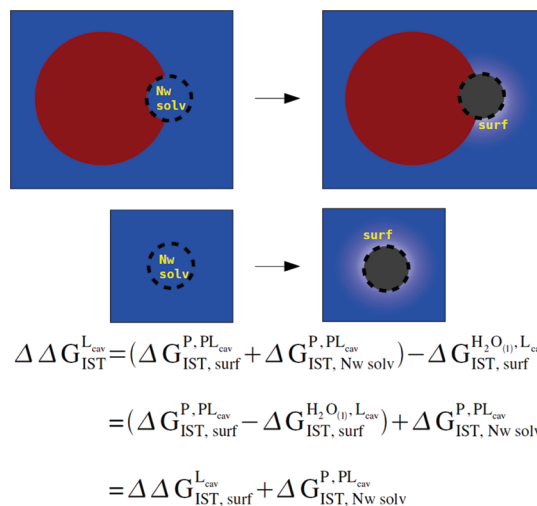


Figure 2. Cartoon depicting the spatial decomposition of the IST integral equations introduced in eqs 11–13. The net “surf” term is the difference in the free energetic cost of the fluid reorganizing its configuration around the surface of the ligand cavity when the cavity is bound to the protein versus free in solution, and the net “N_wsolv” term corresponds to the difference in the *local* IST integral free energy of the N_w water occupying the active site of the protein versus the IST integral free energy of the same N_w water molecules in the bulk fluid.

where the “surf” term is the difference in the free energetic cost of the fluid reorganizing its configuration around the surface of the ligand cavity when the cavity is bound to the protein versus free in solution, and the “N_wsolv” term corresponds to the difference in the *local* IST integral free energy of the N_w water occupying the active site of the protein versus the IST integral free energy of the same N_w water molecules in the bulk fluid. Our final approximation is to assume that for small ligands that are expected to displace only one or a few water molecules deep within the protein active site, the “N_wsolv” term should dominate this expression. Therefore, our final approximation to the binding free energy of the complex is

$$\Delta G_{\text{bind}}^{\circ} \approx \Delta G_{\text{IST},\text{Nw solv}}^{\text{P,PL,cav}} + \delta_{\text{surf}} \Delta\Delta G_{\text{IST},\text{surf}}^{\text{L,cav}} + \delta_{\text{sic}} [\langle U_{\text{P-L}} \rangle_{\text{PL}} + \Delta\langle W_{\text{PL}/\text{P,L;PL}} \rangle^{\text{chrg}}] + \delta_{\text{strn}} [\langle U_{\text{P}} \rangle_{\text{PL}} + \langle U_{\text{L}} \rangle_{\text{PL}} - \langle U_{\text{P}} \rangle_{\text{P}} - \langle U_{\text{L}} \rangle_{\text{L}} - T\Delta S_{\text{config}}^{\circ}] \quad (14)$$

where the differences in the IST “surf” integrals are approximated as negligible when δ_{surf} is set to zero. Thus, our remaining task is to develop a numerical estimate of the “N_wsolv” term.

Interestingly, a possible candidate estimator of $\Delta G_{\text{IST},\text{Nw solv}}^{\text{P,PL,cav}}$ was previously introduced in ref 5, although its connection to the more rigorous expressions for computing protein–ligand binding affinities was not fully clarified at the time of its introduction. In the so-called displaced-solvent functional (DSF) approach, the local values of the IST integrals are computed for regions of high solvent occupancy in the active site, denoted by hydration sites. Note that the volume of each hydration site is chosen such that the number of hydration

sites will correspond to the N_w water molecules that are evacuated from the protein active site to the bulk fluid upon the binding of the ligand. This estimator itself was based on the following assumptions: (1) if atoms of a ligand overlap with a hydration site, they displace the water from that site, and (2) the less energetically or entropically favorable the expelled solvent, the more favorable its contributions to the binding free energy. Thus, the relative binding free energy of the ligand is approximated as

$$\begin{aligned} \Delta G_{\text{IST},N_{\text{wsolv}}}^{\text{P,PL,cav}} &= \Delta G_{\text{bind}}^{\text{DSF}} \\ &= \sum_{\text{lig,hs}} (E_{\text{bulk}} - E_{\text{hs}}) \left(1 - \frac{|\vec{r}_{\text{lig}} - \vec{r}_{\text{hs}}|}{R_{\text{co}}} \right) \Theta(R_{\text{co}} - |\vec{r}_{\text{lig}} - \vec{r}_{\text{hs}}|) \\ &+ T \sum_{\text{lig,hs}} S_{\text{hs}}^{\text{e}} \left(1 - \frac{|\vec{r}_{\text{lig}} - \vec{r}_{\text{hs}}|}{R_{\text{co}}} \right) \Theta(R_{\text{co}} - |\vec{r}_{\text{lig}} - \vec{r}_{\text{hs}}|) \\ &= \sum_{\text{lig,hs}} \Delta G_{\text{hs}} \left(1 - \frac{|\vec{r}_{\text{lig}} - \vec{r}_{\text{hs}}|}{R_{\text{co}}} \right) \Theta(R_{\text{co}} - |\vec{r}_{\text{lig}} - \vec{r}_{\text{hs}}|) \quad (15) \end{aligned}$$

where $\Delta G_{\text{bind}}^{\text{DSF}}$ is the predicted binding free energy of the ligand, R_{co} is the distance cutoff for a ligand atom beginning to displace water from a hydration site, E_{hs} is the system-interaction energy of water in a given hydration site, S_{hs}^{e} is the excess entropy of water in a given hydration site, ΔG_{hs} is the computed free energy of transferring the solvent in a given hydration site from the active site to the bulk fluid, and Θ is the Heaviside step function. We also capped the contribution from each hydration site, such that it would never contribute more than ΔG_{hs} to $\Delta G_{\text{bind}}^{\text{DSF}}$ no matter how many ligand atoms were in close proximity to it. The value R_{co} might be considered a free parameter. However, an approximate value was adopted by noting that the radius of a carbon atom and a water oxygen atom are both approximately 1.4 Å, thus suggesting that contact distances between a water oxygen atom and a ligand carbon atom less than $0.8(1.4 \text{ Å} + 1.4 \text{ Å}) = 2.24 \text{ Å}$ are statistically improbable due to the stiffness of the van der Waals potential. From the preceding approximate theory, we infer that this approach should yield quantitatively accurate predictions of protein–ligand binding free energies versus the FEP reference data when the ligand is complementary to the protein active site and the sum of the reorganization entropies and energies of the protein and the ligand are small compared to the other terms contributing to binding.

Here, however, the preceding theory also suggests an alternative but related approach to adapting the DSF method to compute the binding free energy of a united atom methane molecule to a model hydrophobic enclosure. Here, since the united atom methane molecule is itself simply a sphere that will occupy a known position in the binding site, we may simply collect statistics from the water molecules observed to occupy the volume that will be later occupied by the binding methane. Thus, clustering is unnecessary. From these data, the energetic and entropic properties of the solvating water can be readily obtained via an application of inhomogeneous solvation theory. Last, it would in principle be possible to approximate the binding free energy of the methane molecule via the one-evacuated-site–one-evacuated-water approximation introduced in ref 5. However, we may also identify an approximate scaling that makes use of the

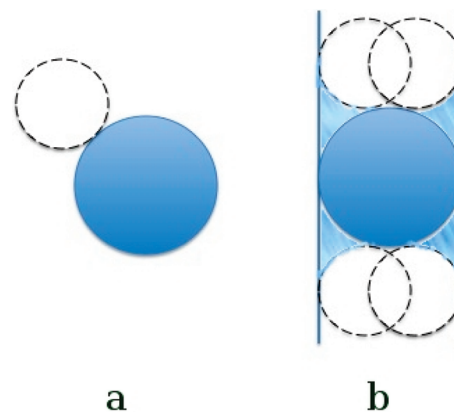


Figure 3. The effective volume displaced by a methane in the bulk (a) and in between two hydrophobic plates (b). The blue particle denotes a methane, and a dashed circle denotes a probe solvent molecule. The volume displaced by a methane in the bulk is just the van der Waals volume of the methane, but in between the two plates, the four corners are also displaced by the methane due to the finite volume of the probe ball.

known volume of the methane particle. In particular, if the methane particle is assumed to have a van der Waals radius of 1.865 Å, then the expectation value of the number of water molecules expected to exist within that volume is

$$N_{\text{eff}} = \rho_{\text{bulk}} \left(\frac{4}{3} \pi R_{\text{methane}}^3 \right) \approx 0.85 \quad (16)$$

where N_{eff} is the effective number of water molecules expected to be displaced by the bound methane assuming the entire system remains at bulk density, Δ_{bulk} is the density of liquid water, and R_{methane} is the van der Waals radius of the methane particle. Clearly, the number density of water in the active site depends on the environment of the specific enclosure and in general would be different from that of the bulk. However, the effective volume that is displaced by the binding methane is also different for different enclosures. Taking the situation of methane between two hydrophobic plates as an example, considering the solvent-excluded volume consisting of the inward-facing surface of the probe ball with a radius of 1.4 Å (size of water), in the bulk water, the volume displaced by methane is just the van der Waals volume of methane, but the four corners are also excluded by the methane in between the two plates (see Figure 3). It is well-known that the number density of water in the hydrophobically enclosed region is smaller than bulk water because of dewetting. Thus, the more enclosed the enclosures are, the smaller the number density of water in the active site, and the larger the effective volume displaced by the methane. These two competing factors make the approximation introduced in eq 16 appropriate for all of the enclosures. In principle, the exact number of excluded water molecules could be identified by the difference in the average number of water molecules surrounding the enclosure in the presence and absence of the bound methane, but this might require excellent statistics to converge.

To numerically test the validity of the preceding theory, we have constructed a series of model hydrophobic encl-

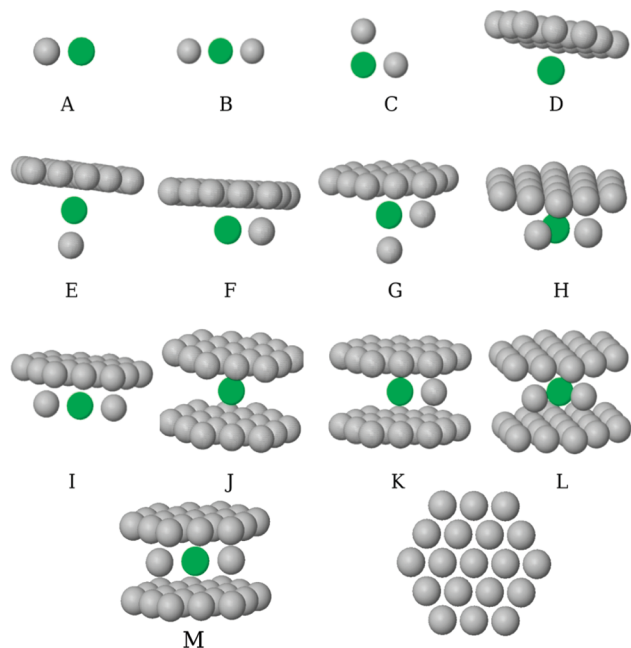


Figure 4. The 13 model hydrophobic enclosures are here depicted in gray. The location of the methane molecule when bound to the respective hydrophobic enclosures is here depicted in green. The geometry of the plate is depicted at the bottom right of this figure. The distance between the neighboring particles in the plate is 3.2 Å, and the distance between the two plates is 7.46 Å. All of the other particles are at contact distance with linear (B, I, and M) and triangle (C, G, H, and L) geometries.

ures, as depicted in Figure 4, and computed the binding free energy of a methane ligand for these hydrophobic enclosures both with FEP theory and the proposed DSF theory. The binding free energies of methane for the described enclosures, as computed by FEP, lie over a 5 kcal/mol range, which would correspond to ~ 4 orders of magnitude of binding affinity. Thus, the ability to accurately predict such free energy differences would be expected to have great utility in a drug-design setting.

A final important point, not relevant to the present model systems but relevant when considering realistic problems such as protein–ligand binding, is the necessity in such real problems for integrating over the solute coordinates. For example, fluctuations of the protein–ligand complex at room temperature can be significant, and in principle this affects the water structure in the active site. In our DSF approach to date, we have employed a single “representative” structure for the protein structure (by harmonically restraining the coordinates to a target structure during the DSF molecular dynamics simulation) rather than allowing the solute phase space to be fully explored. For the model hydrophobic enclosures, there is no issue with averaging over solute configurations because the model enclosures are specified as rigid from the beginning.

In the context of our DSF methodology, the interesting question is how good an approximation the harmonically restrained simulation is to the fully fluctuating solute when estimating the free energy changes resulting from solvent displacement by the ligand. A heuristic argument can be made that the approximation is reasonable if it is assumed

that, for relatively modest fluctuations of the complex (as opposed to major conformational changes), the solvation in the active site “follows” the solute atoms—in essence an adiabatic approximation in which the solvation structure readjusts quickly to typical excursions of solute atoms from the central configuration. If this is in fact the case, then the free energy of displacement of a given water molecule at all accessible solute configurations can be approximated by the displacement free energy at the central configuration. This is not a rigorous or controlled approximation, but it appears to work reasonably well on the basis of a range of examples that we have investigated to date. We do not consider this point further in the present paper, as our focus is on a series of rigid solutes; however, in future work, explicit investigation of this hypothesis, based on computing DSFs for different solute configurations and comparing them, will be pursued.

B. Simulation Details. DSF Analysis. To generate the data required to apply the DSF method of computing protein–ligand binding free energies to the model hydrophobic enclosures, each of the 13 hydrophobic enclosures depicted in Figure 4 were subjected to explicitly solvated molecular dynamics with the Desmond molecular dynamics program.²¹ The Maestro²² System Builder utility was used to insert each enclosure into a cubic water box with a 10 Å buffer. The SPC²³ water model was used to describe the solvent, and the united atom methane molecules that formed the atoms of the enclosures were uniformly represented with $\sigma = 3.73$ Å and $\epsilon = 0.294$ kcal/mol Lennard-Jones parameters. The atoms of the enclosures were constrained to their initial positions throughout their dynamics, and only the solvent degrees of freedom were sampled. The energy of the system was minimized and then equilibrated to 298 K and 1 atm with Nose–Hoover^{24,25} temperature and Martyna–Tobias–Klein pressure²⁶ controls over 500 ps of molecular dynamics. A cutoff distance of 9 Å was used to model the Lennard-Jones interactions, and the particle-mesh Ewald²⁷ method was used to model the electrostatic interactions. Following the equilibration, a 20 ns production molecular dynamics simulation was used to obtain statistics of the water solvating the enclosures, and configurations of the system were collected every 1.002 ps.

Following the previously developed approach,^{5,6} the position the ligand *would occupy* in the enclosures was used to define the active site volume. Here, a 1 Å cutoff distance from the center of where the ligand center would be was used to define the solvent volume of interest. A water molecule was identified to be in the active site when its oxygen lay within the sphere, and otherwise not. For each solvent molecule identified in this volume, we computed the system-interaction energy of the solvent molecule (i.e., the interaction energy of the solvent molecule with the rest of the system) and recorded its orientation and position. From these data, we computed the average system-interaction energy of the solvent occupying this volume, and the excess entropy of this solvent from an expansion of the entropy in terms of translational and orientational correlation functions.

The calculation of excess entropies of water in the hydration sites was processed in a two-step manner: (1)

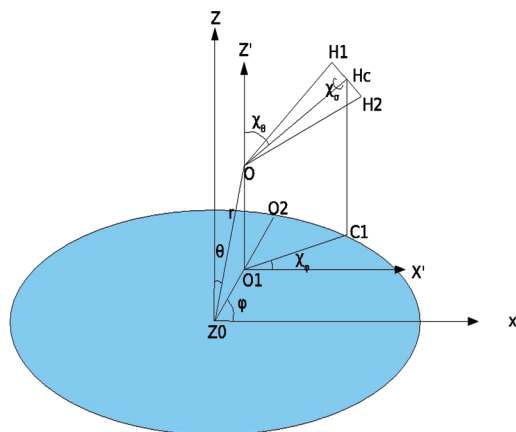


Figure 5. The coordinate system to characterize the position and orientation of water inside the hydration site. The z axis is perpendicular to the model hydrophobic plate, and the x axis is such defined that the other methanes lie on the x axis. r , θ , and φ are the typical spherical coordinates which define the position of the oxygen atom, and χ_θ , χ_φ , and χ_σ are three angles which define the orientation of the water around its oxygen.

introduce an intermediate reference state with the same average number density as the hydration site we are studying but a flat translational and orientational distribution, and calculate the excess entropy of the hydrogen site water with respect to this intermediate reference state due to the local ordering of water in the hydration site; and (2) determine the entropy difference between the intermediate reference state and the bulk water that is due to the difference of number density. The entropy difference between water in the hydration site and the intermediate state was calculated through the integral introduced in eq 12, with $g_{sw}(r)$ defined with respect to the intermediate reference state number density. In order to integrate this entropy expansion, we adopted a k th nearest neighbors approach as introduced in ref 28.

To characterize the orientation of waters in the hydration site, we built the coordinate system such that the center of the hydration site was taken to be the origin, the z axis was perpendicular to the plate (take enclosure F, for example), and a second methane not lying on the z axis was arbitrarily chosen to define the direction of the x axis. The orientation of water in the hydration site was defined by six variables r , θ , φ , χ_θ , χ_φ , and χ_σ , where r , θ , and φ are the typical spherical coordinates which define the position of the oxygen atom and χ_θ , χ_φ , and χ_σ are the three angles which define the orientation of the water around its oxygen (see Figure 5). To clarify, χ_θ and χ_φ are similar to the typical spherical coordinate angles θ and φ , which define the orientation of the dipole vector of water, and χ_σ defines the rotation of hydrogen around the dipole vector. For enclosures with rotational symmetry about the z axis, the distribution along the φ angle is flat by symmetry, so we only need five angles to define the orientation of water. The calculation of the entropy difference is performed through the following equation:

$$S_1 = -k \frac{1}{V\Omega} \int J(r, \theta, \phi, \chi_\theta, \chi_\varphi, \chi_\sigma) g(r, \theta, \phi, \chi_\theta, \chi_\varphi, \chi_\sigma) \times \ln g(r, \theta, \phi, \chi_\theta, \chi_\varphi, \chi_\sigma) dr d\theta d\phi d\chi_\theta d\chi_\varphi d\chi_\sigma \quad (17)$$

where $g(r, \theta, \varphi, \chi_\theta, \chi_\varphi, \chi_\sigma)$ is the solute water pair correlation function (PCF) and $J(r, \theta, \varphi, \chi_\theta, \chi_\varphi, \chi_\sigma)$ is the Jacobian associated with these variables. Here, $g(r, \theta, \varphi, \chi_\theta, \chi_\varphi, \chi_\sigma)$ has the property that

$$\frac{1}{V\Omega} \int J(r, \theta, \phi, \chi_\theta, \chi_\varphi, \chi_\sigma) g(r, \theta, \phi, \chi_\theta, \chi_\varphi, \chi_\sigma) \times dr d\theta d\phi d\chi_\theta d\chi_\varphi d\chi_\sigma = 1 \quad (18)$$

where V is the volume of the sphere and Ω is the total angular volume over angular variables χ_θ , χ_φ , and χ_σ , i.e.,

$$\Omega = \int J(\chi_\theta, \chi_\varphi, \chi_\sigma) g(\chi_\theta, \chi_\varphi, \chi_\sigma) d\chi_\theta d\chi_\varphi d\chi_\sigma \quad (19)$$

In line with ref 28, we approximate the total pair correlation function (PCF) through generalized Kirkwood superposition approximation²⁹ (GKSA), which allowed the entropy to be approximated by the summation and subtraction of one- and two-dimensional entropies, and calculated the one- and two-dimensional entropies through NN method.

The entropy difference between the reference state and bulk water can be simply calculated by recognizing the entropy expression for homogeneous ideal gas:

$$S_{id} = \frac{3}{2} - k \ln(\rho\Lambda^3) \quad (20)$$

where Λ is the thermal wavelength. So the excess entropy of the second step is simply:

$$S_2 = -k \ln\left(\frac{\rho_{ref}}{\rho_{bulk}}\right) \quad (21)$$

where ρ_{ref} and ρ_{bulk} are the number densities of the reference state and bulk water, respectively.

The total excess entropy is the sum of S_1 and S_2 as defined by eqs 17 and 21.

FEP Analysis. The dynamics simulation used to perform the FEP analysis of the binding free energy of the methane ligand to the model hydrophobic enclosures was run under identical simulation protocols as the DSF analysis. The ligand was “turned on” inside the model enclosures over 9 λ windows with $\lambda = [0, 0.125, 0.25, 0.375, 0.50, 0.625, 0.75, 0.875, 1]$, where λ is the coupling parameter to turn on/off the interaction between the methane and the rest of the system with the initial state and final state corresponding to $\lambda = 0$ and $\lambda = 1$, respectively. At different λ windows, we performed molecular dynamics simulations and calculated the energy difference between neighboring λ values for each configuration saved. In these simulations, the soft-core interactions were used for the Lenard-Jones potential. Bennett acceptance ratio methods were then used to calculate the free energy difference between neighboring states. The sum of the free energy differences between neighboring states gave the solvation free energy of methane in question. The same procedure was followed to calculate the solvation free energy of methane in bulk water. The difference between the two solvation free energies gave the binding free energy to bring

Table 1. Binding Thermodynamics of Methane for the Various Model Hydrophobic Enclosures As Computed from DSF Theory and FEP Theory^a

model enclosure ^b	E_{hs} (kcal/mol)	S_{hs}^{e} (cal/mol ^b K)	ΔSASA (\AA^2)	ΔMSA (\AA^2)	ΔE_{LJ} (kcal/mol)	DSF- ΔG_{bind} (kcal/mol)	$N_{\text{eff}}^{\text{p}}$ DSF- ΔG_{bind} (kcal/mol)	FEP- ΔG_{bind} (kcal/mol)
bulk	-19.8	0	0	0	0	0	0	0
A	-19.6	-1.2	-59.45	-3.84	0	-0.5	-0.46	-0.61
B	-18.9	-2.0	-118.9	-7.67	0	-1.5	-1.28	-1.15
C	-19.2	-1.8	-98.21	-10.49	0	-1.1	-0.97	-1.41
D	-18.7	-1.2	-91.32	-13.51	-1.41	-1.5	-1.26	-1.66
E	-17.7	-2.3	-151.15	-17.35	-1.41	-2.8	-2.39	-2.17
F	-17.3	-1.5	-117.52	-24.06	-1.41	-2.9	-2.5	-2.63
G	-16.0	-3.0	-156.39	-30.7	-1.41	-4.7	-4	-3.41
H	-15.6	-1.2	-132.41	-37.35	-1.41	-4.6	-3.92	-3.43
I	-15.6	-1.8	-143.71	-34.6	-1.41	-4.8	-4.05	-3.47
J	-17.8	-2.6	-182.65	-27.02	-2.82	-2.8	-2.41	-2.86
K	-15.5	-2.1	-175.59	-44.27	-2.82	-4.9	-4.17	-4.59
L	-13.0	0.3	-166.61	-64.21	-2.82	-6.8	-5.74	-5.24
M	-13.3	-0.1	-168.52	-61.51	-2.82	-6.6	-5.6	-5.45
R^2 versus FEP	0.94	0.16	0.76 ^c	0.92 ^d	0.73	0.95	0.95	N/A
MAE versus FEP	0.61	N/A	0.54 ^c	0.47 ^d	1.41	0.66	0.36	N/A
RMSE versus FEP	0.75	N/A	0.74 ^c	0.58 ^d	1.63	0.85	0.40	N/A

^a E_{hs} is the hydration site system interaction energy. S_{hs}^{e} is the hydration site solute–water correlation entropy. ΔSASA is the buried solvent accessible surface area using a 1.4 \AA radius probe. ΔE_{LJ} is the Lennard-Jones interaction energy of the bound methane with the rest of the enclosure. DSF- ΔG_{bind} is the predicted binding free energy of the methane molecule for the model enclosure as computed from DSF theory. N_{eff} is scaling coefficient derived by determining the expectation value of the number of water molecules occupying a volume in the bulk fluid equal to the volume of the methane probe molecule, and FEP- ΔG_{bind} is the predicted binding free energy of the methane molecule for the model enclosure as computed from FEP theory. Note that the standard deviation of the E_{hs} values reported below were found to be uniformly less than 0.4 kcal/mol (as obtained from block averaging), and the standard errors of the FEP- ΔG_{bind} values were uniformly less than 0.02 kcal/mol. ^b The enclosures are labeled as described in Figure 3. ^c These values correspond to the correlation between the buried SASA/LJ interaction with an optimized surface tension coefficient ($\gamma = 0.044$ kcal/mol $\cdot\text{\AA}^2$) and the FEP reference data. ^d These values correspond to the correlation between the buried MSA/LJ interaction with an optimized surface tension coefficient ($\delta = 0.011$ kcal/mol $\cdot\text{\AA}^2$) and the FEP reference data.

a methane from infinitely far to inside the hydrophobic enclosure. (We can also interpret the binding free energy as the potential of mean force between the methane and the enclosure.)

Buried Surface Area Analysis. The solvent accessible surface area (SASA) and molecular surface area (MSA, or Connolly surface) of each enclosure with and without the bound methane was computed with the Connolly molecular surface package,³⁰ as was the SASA and MSA of the methane particle by itself. From these data, the buried solvent accessible surface area upon methane-enclosure complexation was determined. The Lennard-Jones interaction energy of the methane particle with the model enclosure was similarly computed. The buried surface area times the surface tension would give the solvent induced interaction energy, and together with the direct Lennard-Jones interaction energy, the total binding energy of methane with different enclosures can be calculated, as routinely estimated in various empirical methods to estimate the contribution of the nonpolar term to the binding energy.

Results

The binding free energies of methane for the model hydrophobic enclosures, as measured by FEP, are reported in Table 1. It is found that the range of binding free energies of the methane ligand for the model enclosures is nearly 5 kcal/mol. Also reported in Table 1 are the system-interaction energies and excess entropies of the water displaced by the methane ligand, the buried surface area upon complexation (both SASA and MSA), the change of the Lennard-Jones interaction energy between the methane particle and the

enclosure upon complexation, the DSF prediction of the binding free energy of the complex, and the scaled DSF prediction that makes use of the scaling coefficient deduced from first principles in section II. The R^2 value, mean absolute error (MAE), and the root-mean-square error (RMSE) between the various predictions with the FEP reference data are also listed in the last few rows of the table. Note here that the surface tension coefficients for the buried surface area/molecular mechanics predictions (both SASA and MSA) were explicitly tuned to minimize the MAE of the predictions. Such explicit tuning yields significantly better results than could reasonably be expected to be obtained if such methods were employed with fixed coefficients across realistically variable data sets.

The DSF predictions show very high correlation with the FEP reference data, as indicated by the R^2 value of 0.95 (Figure 6), where the buried surface area/Lennard-Jones interaction predictions show reduced correlations, as indicated by R^2 values of 0.92 for MSA/MM and 0.76 for SASA/MM (Figure 7). The DSF method also allows for the decomposition of the binding free energy prediction into separate enthalpic and entropic components. Inspection of the data reported in Table 1 indicates that the DSF predictions are dominated by the enthalpic contribution to the binding affinity, which by itself manifests an R^2 value of 0.94 versus the FEP reference data. Detailed analysis of these data indicates that, except for the first three systems, the binding of the methane molecule to these hydrophobic enclosures is mainly an enthalpy driven event, which is consistent with our knowledge about large length scale hydrophobicity.^{31–33} Recent calorimetry data obtained for Major Mouse Urinary

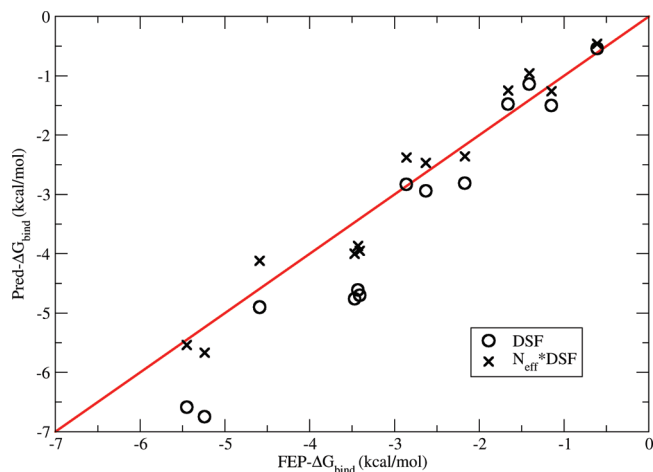


Figure 6. The correlation of the DSF predictions of the methane-enclosure binding free energies with the FEP reference data.

Protein by Homans et al.³⁴ appear to indicate that such enthalpy-driven hydrophobic binding events are witnessed in vivo, as well.

The inspection of the trajectory indicates that the atomistic basis of the enthalpy driven effect is that water molecules that solvate such enclosures are forced to break hydrogen bonds. The effect is most obvious for hydrophobic enclosures L and M, where the solvent suffers a ~ 7 kcal/mol reduction in system-interaction energy when occupying these enclosures, while there is almost no reduction in excess entropy versus bulk water. Conversely, the methane dimerization free energy described by methane binding to “enclosure” A is dominated by the entropic contribution, again consistent with an entropy-driven small length scale hydrophobic effect. This finding is analogous to the well characterized length scale dependence of the hydrophobic effect; while small hydrophobes are found to induce entropic ordering of the solvent, large hydrophobes are found to break water–water hydrogen bonds.^{31,33} The enclosures L and M can thus be understood as manifesting extreme large-length-scale hydrophobic character from the perspective of the solvating water.

Figure 6 plots the correlation of the DSF binding free energies versus the FEP reference data with and without the derived scaling coefficient deduced from the size of the methane ligand itself. As can be seen from the figure, both sets of predictions track the FEP reference data quite well. However, the scaled predictions have greater quantitative agreement with the FEP, which may be quantified by the mean absolute error (MAE) and root-mean-square error (RMSE) metrics. Here, the scaled predictions are found to have a MAE of 0.36 kcal/mol and a RMSE of 0.40 kcal/mol, while the unscaled predictions have a MAE of 0.66 kcal/mol and a RMSE of 0.84 kcal/mol. Thus, the deduced scaling coefficient appears to increase the quantitative accuracy of the approach, in line with the expectation of the theoretical analysis.

We also investigated to what extent a combined buried surface area/Lennard-Jones interaction energy model might be able to reproduce the binding affinities. Tuning the model to minimize the MAE of the fit, we obtained an optimal surface tension coefficient of $\gamma = 0.011$ kcal/mol $\cdot\text{\AA}^2$ for

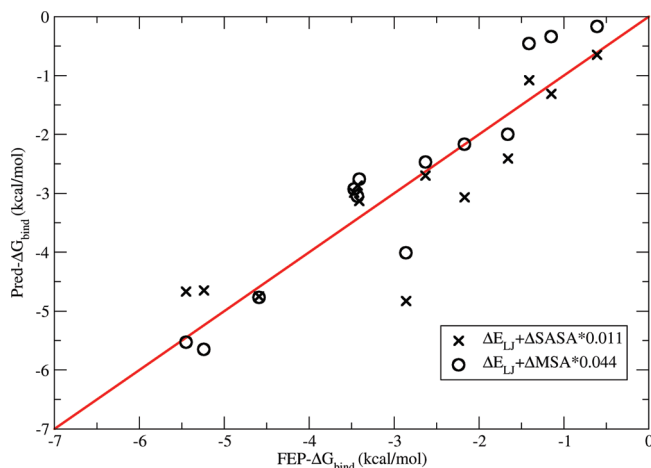


Figure 7. The correlation of buried surface area/molecular mechanics predictions of the methane-enclosure binding free energies with the FEP reference data. The water SASA surface tension coefficient (0.011 kcal/mol $\cdot\text{\AA}^2$) and MSA surface tension coefficient (0.044 kcal/mol $\cdot\text{\AA}^2$) were tuned to minimize the absolute average error of the predictions with respect to the reference data.

SASA and 0.044 kcal/mol $\cdot\text{\AA}^2$ for MSA for these enclosures, which is somewhat smaller than the reported literature values.³⁵ These predictions versus the FEP reference data are reported in Figure 7. It is found that MSA/MM performed much better compared with SASA/MM, which is indicated by a much higher R^2 value and smaller MAE and RMSE values (data listed in last three rows in Table 1). However, both of them performed less well than the DSF predictions with the scaling coefficient correction, and much worse results would be expected with such a model in general, as noted above, since it would not benefit from explicit fitting to the reference data.

The better performance of MSA/MM versus SASA/MM is due to the better characterization of MSA for the topology of enclosures J, K, L, and M. SASA/MM predicts enclosure J to be the most hydrophobic, which corresponds to a methane molecule binding between two hydrophobic plates, because large swaths of formerly SASA on the faces of the plates are buried by the presence of the methane ligand for enclosure J, while for enclosures K, L, and M several methane molecules already lie between the plates in the absence of the binding ligand, and thus some of the surface area that would be buried by the binding methane is already buried by the other particles. However, MSA can better characterize the curvature of these enclosures and predict the correct rank order of the binding affinities.

Conclusion

Calculations suggest that the DSF method of computing protein–ligand binding affinities may offer near-FEP accuracy at a substantially reduced computational expense for systems that satisfy the requisite approximations and should offer greater quantitative accuracy than competing implicit solvent methodologies. Further, the clear connection between the DSF method and more rigorous statistical mechanical expressions may offer a rational path to systematically

improve the accuracy and rigor of the method by progressive inclusion of those counter-balancing terms currently approximated to exactly cancel. This previously unclarified connection to the underlying theory facilitated the derivation of a scaling coefficient that was seen to increase the quality of the predictions of the method versus the FEP reference data. Last, the molecular detail afforded by the technique may offer insight into protein–ligand binding processes, such as highlighting the importance of the enthalpy in the binding of methane to such model enclosures, which may have been difficult to discern from only FEP or implicit modeling.

Acknowledgment. This work was supported by NIH grants to B.J.B. (NIH GM 43340) and to R.A.F. (NIH GM 52018), and an NSF fellowship to R.A. B.J.B. and R.A.F. acknowledge that this work was also supported in part by the National Science Foundation through TeraGrid resources provided by NCSA and ABE (MCA08X002).

References

- (1) Gilson, M. K.; Zhou, H. X. Calculation of protein-ligand binding affinities. *Annu. Rev. Biophys. Biomol. Struct.* **2007**, *36*, 21–42.
- (2) Mobley, D. L.; Dill, K. A. Binding of Small-Molecule Ligands to Proteins: “What You See” Is Not Always “What You Get”. *Structure*. **2009**, *17*, 489–98.
- (3) Zhou, H. X.; Gilson, M. K. Theory of Free Energy and Entropy in Noncovalent Binding. *Chem. Rev.* **2009**, *109*, 4092–4107.
- (4) Guvench, O.; MacKerell, A. D. Computational evaluation of protein-small molecule binding. *Curr. Opin. Struct. Biol.* **2009**, *19*, 56–61.
- (5) Abel, R.; Young, T.; Farid, R.; Berne, B. J.; Friesner, R. A. Role of the active-site solvent in the thermodynamics of factor Xa ligand binding. *J. Am. Chem. Soc.* **2008**, *130*, 2817–2831.
- (6) Young, T.; Abel, R.; Kim, B.; Berne, B. J.; Friesner, R. A. Motifs for molecular recognition exploiting hydrophobic enclosure in protein-ligand binding. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104*, 808–813.
- (7) Clausen, R. P.; Naur, P.; Kristensen, A. S.; Greenwood, J. R.; Strange, M.; Brauner-Osborne, H.; Jensen, A. A.; Nielsen, A. S. T.; Geneser, U.; Ringgaard, L. M.; Nielsen, B.; Pickering, D. S.; Brehm, L.; Gajhede, M.; Krosgaard-Larsen, P.; Kastrup, J. S. The Glutamate Receptor GluR5 Agonist (S)-2-Amino-3-(3-hydroxy-7,8-dihydro-6H-cyclohepta[d]isoxazol-4-yl)propionic Acid and the 8-Methyl Analogue: Synthesis, Molecular Pharmacology, and Biostructural Characterization. *J. Med. Chem.* **2009**, *52*, 4911–4922.
- (8) Beuming, T.; Farid, R.; Sherman, W. High-energy water sites determine peptide binding affinity and specificity of PDZ domains. *Protein Sci.* **2009**, *18*, 1609–1619.
- (9) Robinson, D. D.; Sherman, W.; Farid, R. Understanding Kinase Selectivity Through Energetic Analysis of Binding Site Waters. *ChemMedChem* **2010**, *5*, 618–627.
- (10) Guimaraes, C. R. W.; Mathiowetz, A. M. Addressing Limitations with the MM-GB/SA Scoring Procedure using the Water Map Method and Free Energy Perturbation Calculations. *J. Chem. Inf. Model.* **2010**, *50*, 547–559.
- (11) Pearlstein R. A.; Hu Q. Y.; Zhou J.; Yowe D.; Levell J.; Dale B.; Kaushik V. K.; Daniels D.; Hanrahan S.; Sherman W.; Abel R. New hypotheses about the structure-function of proprotein convertase subtilisin/kexin type 9: Analysis of the epidermal growth factor-like repeat A docking site using WaterMap. *Proteins* **2010**, in press. DOI: 10.1002/prot.22767.
- (12) Chrencik J. E.; Patny A.; Leung I. K.; Korniski B.; Emmons T. L.; Hall T.; Weinberg R. A.; Gormley J. A.; Williams J. M.; Day J. E.; Hirsch J. L.; Kiefer J. R.; Leone J. W.; Fischer H. D.; Sommers C. D.; Huang H. C.; Jacobsen E. J.; Tenbrink R. E.; Tomasselli A. G.; Benson T. E. Structural and Thermodynamic Characterization of the TYK2 and JAK3 Kinase Domains in Complex with CP-690550 and CMP-6. *J. Mol. Biol.* **2010**, in press. DOI: 10.1016/j.jmb.2010.05.020.
- (13) Mobley, D. L.; Chodera, J. D.; Dill, K. A. Confine-and-release method: Obtaining correct binding free energies in the presence of protein conformational change. *J. Chem. Theory Comput.* **2007**, *3*, 1231–1235.
- (14) Deng, Y. Q.; Roux, B. Computations of Standard Binding Free Energies with Molecular Dynamics Simulations. *J. Phys. Chem. B.* **2009**, *113*, 2234–2246.
- (15) Swanson, J. M. J.; Henchman, R. H.; McCammon, J. A. Revisiting free energy calculations: A theoretical connection to MM/PBSA and direct calculation of the association free energy. *Biophys. J.* **2004**, *86*, 67–74.
- (16) Huang, N.; Kalyanaraman, C.; Bernacki, K.; Jacobson, M. P. Molecular mechanics methods for predicting protein-ligand binding. *Phys. Chem. Chem. Phys.* **2006**, *44*, 5166–5177.
- (17) Guimarães, C. R.; Cardozo, M. MM-GB/SA rescoring of docking poses in structure-based lead optimization. *J. Chem. Inf. Model.* **2008**, *48*, 958–970.
- (18) Mobley, D. L.; Bayly, C. I.; Cooper, M. D.; Shirts, M. R.; Dill, K. A. Small Molecule Hydration Free Energies in Explicit Solvent: An Extensive Test of Fixed-Charge Atomistic Simulations. *J. Chem. Theory Comput.* **2009**, *5*, 350–358.
- (19) Gallicchio, E.; Kubo, M. M.; Levy, R. M. Enthalpy-entropy and cavity decomposition of alkane hydration free energies: Numerical results and implications for theories of hydrophobic solvation. *J. Phys. Chem. B* **2000**, *104*, 6271–6285.
- (20) Lazaridis, T. Inhomogeneous fluid approach to solvation thermodynamics. 1. Theory. *J. Phys. Chem. B.* **1998**, *102*, 3531–3541.
- (21) Bower, K. L. *SC2006 November 2006*; 0-7695-2700-0/062006, IEEE: Tampa, FL.
- (22) Banks, J. L.; Beard, H. S.; Cao, Y. X.; Cho, A. E.; Damm, W.; Farid, R.; Felts, A. K.; Halgren, T. A.; Mainz, D. T.; Maple, J. R.; Murphy, R.; Philipp, D. M.; Repasky, M. P.; Zhang, L. Y.; Berne, B. J.; Friesner, R. A.; Gallicchio, E.; Levy, R. M. *J. Comput. Chem.* **2005**, *26*, 1752–1780.
- (23) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Hermans, J. *Intermol. Forces.* **1981**, 331–342.
- (24) Nose, S. A unified formulation of the constant temperature molecular-dynamics methods. *J. Chem. Phys.* **1984**, *81*, 511–519.
- (25) Hoover, W. G. Canonical dynamics - equilibrium phase-space distributions. *Phys. Rev. A* **1985**, *31*, 1695–1697.
- (26) Martyna, G. J.; Tobias, D. J.; Klein, M. L. Constant-pressure molecular-dynamics algorithms. *J. Chem. Phys.* **1994**, *101*, 4177–4189.
- (27) Darden, T.; York, D.; Pedersen, L. Particle mesh Ewald - An N.LOG(N) method for Ewald sums in large systems. *J. Chem. Phys.* **1993**, *98*, 10089–10092.

- (28) Wang, L.; Abel, R.; Friesner, R. A.; Berne, B. J. Thermodynamic Properties of Liquid Water: An Application of a Nonparametric Approach to Computing the Entropy of a Neat Fluid. *J. Chem. Theory Comput.* **2009**, *5*, 1462–1473.
- (29) Singer, A. Maximum entropy formulation of the Kirkwood superposition approximation. *J. Chem. Phys.* **2004**, *121*, 3657–66.
- (30) Connolly, M. L. The molecular surface package. *J. Mol. Graphics* **1993**, *11*, 139–141.
- (31) Berne, B. J.; Weeks, J. D.; Zhou, R. Dewetting and Hydrophobic Interaction in Physical and Biological Systems. *Annu. Rev. Phys. Chem.* **2009**, *60*, 85–103.
- (32) Hummer, G.; Garde, S.; Garcia, A. E.; Paulaitis, M. E.; Pratt, L. R. Hydrophobic effects on a molecular scale. *J. Phys. Chem. B.* **1998**, *102*, 10469–10482.
- (33) Southall, N. T.; Dill, K. A.; Haymet, A. D. J. A view of the hydrophobic effect. *J. Phys. Chem. B.* **2002**, *106*, 521–533.
- (34) Homans, S. W. Water, water everywhere - except where it matters? *Drug Discovery Today* **2007**, *12*, 534–539.
- (35) Sharp, K. A.; Nicholls, A.; Fine, R. F.; Honig, B. Reconciling the magnitude of the microscopic hydrophobic hydrophobic effects. *Science* **1991**, *252*, 106–109.

CT100215C

JCTC

Journal of Chemical Theory and Computation

Structural Survey of Zinc-Containing Proteins and Development of the Zinc AMBER Force Field (ZAFF)

Martin B. Peters, Yue Yang, Bing Wang, László Füsti-Molnár, Michael N. Weaver, and Kenneth M. Merz, Jr.*

Department of Chemistry, Quantum Theory Project, 2328 New Physics Building, P.O. Box 118435, University of Florida, Gainesville, Florida 32611-8435

Received May 18, 2010

Abstract: Currently the Protein Data Bank (PDB) contains over 25 000 structures that contain a metal ion including Na, Mg, K, Ca, V, Cr, Mn, Fe, Co, Ni, Cu, Zn, Pd, Ag, Cd, Ir, Pt, Au, and Hg. In general, carrying out classical molecular dynamics (MD) simulations of metalloproteins is a convoluted and time-consuming process. Herein, we describe MCPB (Metal Center Parameter Builder), which allows one to conveniently and rapidly incorporate metal ions using the bonded plus electrostatics model (Hoops et al. *J. Am. Chem. Soc.* 1991, **113**, 8262–8270) into the AMBER force field (FF). MCPB was used to develop a zinc FF, ZAFF, which is compatible with the existing AMBER FFs. The PDB was mined for all Zn-containing structures with most being tetrahedrally bound. The most abundant primary shell ligand combinations were extracted, and FFs were created. These include Zn bound to CCCC, CCCH, CCHH, CHHH, HHHH, HHHO, HHOO, HOOO, HHHD, and HHDD (O = water and the remaining are 1-letter amino acid codes). Bond and angle force constants and RESP charges were obtained from B3LYP/6-31G* calculations of model structures from the various primary shell combinations. MCPB and ZAFF can be used to create FFs for MD simulations of metalloproteins to study enzyme catalysis, drug design, and metalloprotein crystal refinement.

Introduction

There are currently over 66 000 structures in the Protein Data Bank (PDB), and searching for “metal” results in over 25 000 hits with an approximate break down shown in Table 1. Metal ions play a vital role in protein function, structure, and stability, with zinc, iron, manganese, and copper transition metals being well represented in the PDB (see Table 1). It is desirable to model metalloprotein systems using MM models because one can carry out simulations to address important structure/function and dynamics questions that are not currently attainable using QM and/or QM/MM-based methods due to unavailability of parameters or simply system size.

There are a number of approaches to incorporating metal ions into FFs. The bonded model defines bonds, angles, and torsions between the metal ion and its ligand, which are

added to the FF plus the van der Waals component of the nonbonded function. Hancock^{1,2} used this approach to study systems including copper and nickel. The bonded plus electrostatics model³ defines bonds and angles between the metal ion and its ligand as well as electrostatic potential (ESP) charges (Figure 1a). This method attempts to define the correct electrostatic representation of the metal active site because simply assigning a plus two formal charge to a

Table 1. Metals in the Protein Data Bank (July 2010)

metal	hits	metal	hits	metal	hits
Na	2701	V	59	Pd	8
Mg	5384	Cr	7	Ag	10
K	965	Mn	1412	Cd	521
Ca	5030	Fe	1403	Ir	2
		Co	337	Pt	44
		Ni	463	Au	30
		Cu	763	Hg	343
		Zn	5854		
		total	= 25 336		

* Corresponding author phone: 352-392-6973; fax: 352-392-8722; e-mail: merz@qtp.ufl.edu.

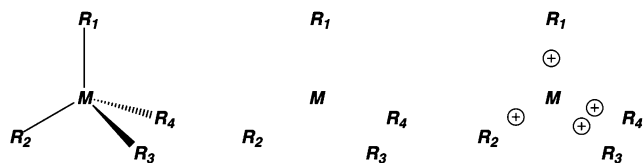


Figure 1. Three approaches to incorporate metal atoms into molecular mechanics force fields. The bonded model (left) defines bonds, angles, and dihedrals between the metal and the ligands, while the nonbonded model (middle) does not and uses electrostatics and van der Waals to model the interactions. The cationic dummy atom model (right) is a derivative of the nonbonded model where cations are placed near the metal center to mimic valence electrons around the metal.

divalent metal ion would not describe the reality of the electronic structure of a metal/ligand complex even though formally correct. The partial atomic charges can be calculated using the RESP approach⁴ or the CMX models of Truhlar and Cramer.⁵ The bond and angle force constants are derived from experiment or calculated using *ab initio* or DFT methods, while the torsion term has so far been neglected. The nonbonded model does not define any extra bonds and places integer charge on the metal ion.⁶ Electrostatic and Lennard–Jones terms describe the interactions. Modifications to this model to include polarization and charge transfer effects have been developed (Figure 1b).⁷ The cationic dummy atom model is related to the nonbonded method where it places dummy atoms (cations) to mimic valence electrons around the metal ion.⁸ Electrostatic and Lennard–Jones terms between the dummy atoms and ligating residues describe the metal–ion interactions (Figure 1c).^{9,10}

Other methods include those of Vedani et al., which is a compromise between the bonded and the nonbonded methods and is implemented in the YETI program,^{11,12} the SIBFA of Gresh and co-workers,^{13,14} and the universal force field (UFF) of Goddard and Rappe and co-workers.^{15–17} These methods do not use a pairwise additive potential or are not readily available in typical biomolecular modeling packages.

Carrying out MM modeling or MD simulations of metal-containing proteins is a complicated procedure using the bonded plus electrostatics model. Incorporating metals into protein force fields is a convoluted process due to the plethora of QM Hamiltonians, basis sets, and charge models to choose from. Also, it has generally been carried out by hand without extensive validation for specific metalloproteins. Some of the published force fields for zinc-, copper-, nickel-, iron-, and platinum-containing systems using the bonded plus electrostatics model are listed in Table 2. There have been numerous other FFs containing various metals published including ruthenium(II)–polypyridyl,¹⁸ cobalt corrinoids,^{19–22} Staphylococcal Nuclease,²³ alcohol dehydrogenase,^{24,25} and metalloporphyrins.^{26–30}

Automated procedures for the parametrization of MM functions for inorganic coordination chemistry have been developed over the last number of years by Norrby and co-workers.^{31,32} Their attempts have focused on generating parameters using experimental, structural data from databases such as the Cambridge Structural Database (CSD) and quantum mechanical reference data using a version of the

Table 2. Published Metalloprotein Force Fields Using the Bonded Plus Electrostatics Model

metal	protein/DNA
zinc	human carbonic anhydrase ^{3,36,37} beta-lactamase ³⁸ dinuclear beta-lactamase ^{39,40} farnesyl transferase ⁴¹
copper	blue copper proteins ^{42–46}
nickel	urea amidohydrolase ^{47,48} NikR ³⁴
iron	cytochrome P450 ^{49,50}
platinum	DNA/cisplatin ⁵¹
copper, zinc	superoxide dismutase ⁵²

MM3 force field.³³ Recently, we used the approach described herein to build a FF for the Ni-containing metal center in NikR,³⁴ but the full computational details were not presented. Herein, we present these details in the context of a Zn FF. We note that a similar approach was published after our NikR work by Lin and Wang,³⁵ but a program or framework with which people can easily and consistently build parameters for metal-containing protein systems was not fully developed. The advantage of a program, like the one reported here, is it is possible to quickly build a metal ion FF for the typically troublesome metal sites that are common in naturally occurring biological systems.

Herein we describe an approach to rapidly build, prototype, and validate MM-based FFs for metalloprotein systems using the AMBER FF. The base approach will be the bonded plus electrostatics model, as we have the most experience and are most comfortable with this approach. However, the tools created can be readily utilized to develop another class of metal ion FF models including nonbonded models.

Implementation. The goal of this research was to provide a platform to rapidly build, prototype, and validate MM models of metalloproteins using the bonded plus electrostatics model for the AMBER suite of programs.⁵³ The bonded plus electrostatics model was chosen over the other approaches as the resulting parameters fit the FF functional forms used in AMBER⁵⁴ and CHARMM,⁵⁵ which are two widely used biomolecular simulation packages. Additionally, the functions used in these programs are pairwise additive, meaning there are no cross-terms and are thus easier to parametrize and less computationally expensive. The latter is a key point when considering that fully solvated metalloproteins in MD simulations can have many hundreds of thousands of atoms. A computer program MCPB (Metal Center Parameter Builder) to generate FF parameters for metalloproteins was developed to this end. MCPB was not built to supersede the approaches developed by Norrby described above but instead to incorporate a realistic bonded and electrostatic model of the metal center into the AMBER FF. The MCPB program was built using the MTK++ Application Program Interface (API), which was developed as an in-house modeling platform for metalloproteins. A complete workflow of MCPB can be seen in Figure 2. The MCPB program carries out the following steps after a structure is downloaded from the Protein Data Bank (PDB). First, the program checks whether the structure contains a transition metal. If the structure does not contain a metal

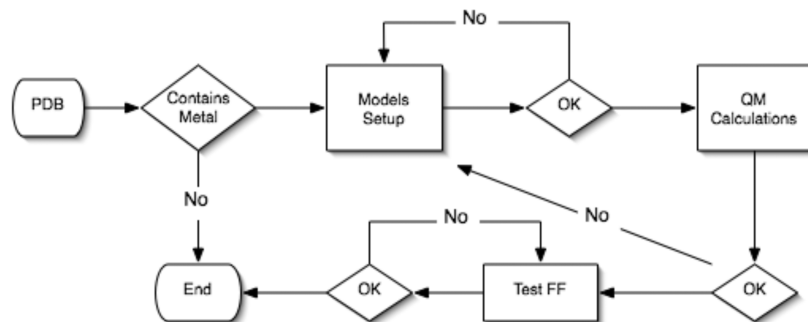


Figure 2. MCPB flow diagram.

then the program terminates. Otherwise, MCPB attempts to determine the primary and secondary ligands of the metal using rules described by Harding,^{56–61} which will be described in more detail below. Once a metal site is found, MCPB creates model structures of the metal's first coordination sphere with which ab initio calculations can be performed on to generate AMBER-like FF parameters. These models include one to generate charges, q_i , and another to determine bond, K_r , and angle, K_θ , force constants. The AMBER function includes bond, angle, torsion, improper, van der Waals, and electrostatic terms; however, only bond, angle, and electrostatic terms are parametrized under the assumption made by Hoops et al. that dihedral terms can be ignored.³ Lennard–Jones parameters are also not parametrized here due to the fact that most metals are buried and that van der Waals interactions are not as important as the electrostatics.³² Lennard–Jones parameters for the most common metal ion in biology were taken from the literature.^{62–66} The methods of incorporating the bond, angle, and charge parameters are outlined below. Once a FF is produced it is tested using minimization techniques to observe its stability. Further validation tools such as comparing the frequencies from both ab initio and the resulting FF could be used as well.⁶⁷

Equilibrium Bond Lengths and Angles. Equilibrium values for bonds, r_{eq} , and angles, θ_{eq} , can be determined through ab initio calculations or taken directly from the crystal structure in the PDB. There are pros and cons for using values from both methods. Ab initio calculations are generally carried out in the gas phase, but solvent effects can be incorporated with, for example, PCM but with an added cost. Crystal structures may contain spurious values and may not be representative of all structures with this bond or angle type. Therefore, the values from ab initio calculations were used here. Alternatively, data from the CSD could be utilized as well, but as with the approaches described above this method has pros and cons as well. Given our focus on metalloproteins and not solid-state metal clusters we have not compared with or used CSD-derived information in the present article.

Force Constants. The force constants, K_r , and K_θ , were obtained by first creating a model (model 1) of the metal site, adding hydrogen atoms using the functionality within MTK++ and then optimizing it in the gas phase. The residues bound to the metal were approximated, for example, cysteine by a thiolate or histidine by a methyl-imidazole, to reduce the computational cost of the minimization. However,

all bonds and angles missing from the FF were accounted for. Once a minimum was found the second derivatives were determined. The Cartesian Hessian matrix is shown in eq 1, which is the second derivative of energy with respect to coordinates. The eigen analysis of k provides the force constants, λ_i , and the normal modes, \hat{v}_i , as shown in eq 2. The interatomic force constant, K_{AB} , between atoms A and B is required to determine the force on atom A by displacing atom B as shown in eq 3 which is required for a MM function.

$$[k] = k_{ij} = \frac{\partial^2 E}{\partial x_i \partial x_j} \quad (1)$$

$$F_i = -[k]\hat{v}_i \delta d = -\lambda_i \hat{v}_i \delta r \quad (2)$$

$$\delta F_A = [k_{AB}]\delta r_B \quad (3)$$

From the minimized structure of model 1 the metal–ligand bond and angle force constants were evaluated. The force constants were converted from Cartesian into internal coordinates using the Gaussian program⁶⁸ providing the following keyword (iop(7/33 = 1)). Force constants were also determined using a method described by Seminario,⁶⁹ and mathematical details can be found in the Supporting Information. Briefly, the force constants are calculated from submatrices of the Cartesian Hessian matrix. This method has the advantage over the “traditional” method of determining force constants as it avoids defining internal coordinates. The MCPB program then reads either the internal force constant matrix or the Seminario-derived parameters and assigns the values to the appropriate bonds and angles.

Point Charges. The atom-centered partial charges were derived using the Merz–Singh–Kollman (MK)⁷⁰ and restrained electrostatic potential (RESP)^{4,71,72} schemes using a second model (model 2) of the metal center. This model included all atoms of a bound residue, which were capped with acetyl (ACE) and *N*-methylamine (NME) residues. If two ligating residues were less than five residues apart then they were tethered with glycine residues and the chain capped with ACE and NME. Again, hydrogen atoms were added using MTK++. This model was not allowed to relax to save computational expense and to keep the crystallographic geometry. The van der Waals radii for the metals used in the MK scheme were taken from the literature.³ The MK/RESP scheme was favored over other charge model schemes because its ability to adjust the charge of the capped or

Table 3. Metal–Donor Bond Target Lengths (in Ångstroms)^a

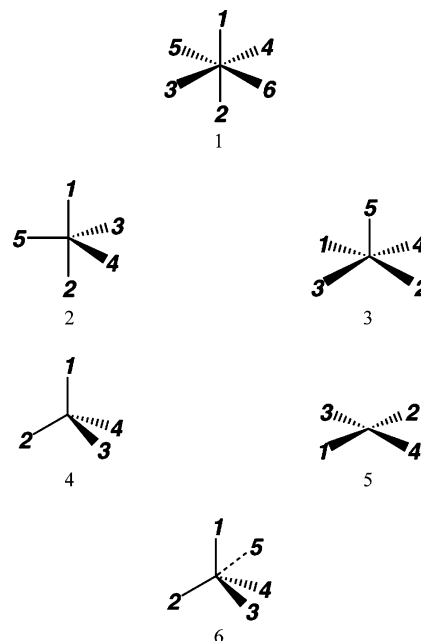
metal	HOH	ASP/GLU	HIS	CYS/MET	SER/THR	TYR	CRL
Na	2.41	2.41					2.38
Mg	2.07	2.07			2.10	1.87	2.26
K	2.81	2.82					2.74
Ca	2.39	2.36			2.43	2.20	2.36
Mn	2.19	2.15	2.21	2.35	2.25	1.88	2.19
Fe	2.09	2.04	2.16	2.30	2.13	1.93	2.04
Co	2.09	2.05	2.14	2.25	2.10	1.90	2.08
Ni	2.09	2.05	2.14	2.25	2.10	1.90	2.08
Cu	2.13	1.99	2.02	2.15	2.00	1.90	2.04
Zn	2.09	1.99	2.03	2.31	2.14	1.95	2.07

^a The following donor atoms of residues are implied: HOH@O, ASP@OD1/OD2, GLU@OE1/OE2, HIS@ND1/NE2, CYS@SG, MET@SG, SER@O, THR@O, TYR@O, and the amino acid backbone carbonyl oxygen atom CRL. If a metal–donor distance is within these target distances plus some tolerance (0.5 Å) it is considered a primary interaction.

linking residues to an integer value, thus allowing the formal charge of the cluster to disperse over the metal and the bound ligands.

In this work, four different methods to develop charges were implemented. The first method allows all atoms of the bound residue to change (ChgModA), the second technique restrains the backbone heavy atoms (CA, N, C, O) to those values found in the AMBER parm94 force field (ChgModB), the third one restrains all the backbone atoms (CA, H, HA, N, HN, C, O) to the AMBER parm94 force field values (ChgModC), while the fourth one (ChgModD) also adds carbon beta (CB) into the restraint list.

Zinc AMBER Force Field. With the ability to build metal FFs established, the task of generating a FF that facilitates simulations for the majority of Zn proteins was initiated. Zinc was chosen because a considerable number of proteins contain that metal as highlighted in Table 1 while also being computationally well behaved. Metalloproteins containing zinc are both structural and functional proteins, and in general, Zn is four coordinate, although sometimes five or six coordinate when multiple ASP/GLU residues or water molecules bind. It was then necessary to determine all Zn environments that exist in proteins. This was carried out using a program called pdbSearcher to analyze all structures currently in the PDB. Again, pdbSearcher was developed using the API provided by MTK++. All X-ray crystal structures with a resolution below 3.0 Å were extracted from a local mirror of the PDB for further analysis. For each metal site the primary and secondary shell ligands were determined using Harding's bond cutoff values as shown in Table 3.^{56–61} These values were determined from a series of papers describing metal coordination in the CSD. A donor atom is considered to be in the primary coordination shell of a metal ion if it is within the target distance as shown in Table 3 plus some tolerance (0.5 Å was used). Metal–donor distances lying between the target distance plus the tolerance and the target distance plus a second tolerance (1.0 Å was used) were defined as secondary shell ligands. For example, if a Zn atom is less than 2.53 Å from a Histidine ND1 or NE2 atom then it is considered a primary ligand. If it was less than 3.03 Å away then that ligand is labeled as secondary; otherwise, it is unbound. Once the number of primary and secondary shell

**Figure 3.** Metal–ligand geometries perceived using Harding's rules.**Table 4.** Ideal Angles Used To Calculate Root Mean Square Deviations for Tetrahedral, Square Planar, Trigonal Bipyramidal, Square Pyramid, and Octahedral Geometries^a

type	coordination	angle	atoms
ML ₄	tetrahedral	109.5°	
	square planar	180.0°	a ₁₂ , a ₃₄
		90.0°	all others
ML ₅	trigonal bipyramidal	180.0°	a ₁₂
		120.0°	a ₃₄ , a ₄₅ , a ₃₅
	square pyramid	90.0°	all others
		$b_m = (1)/(4) \sum_{i=1}^4 1/a_{i5}$ $(360.0 - 2b_m)$ $2 \sin^{-1}(2^{-1/2} \sin(180.0 - b_m))$	a ₁₅ , a ₂₅ , a ₃₅ , a ₄₅ a ₁₂ , a ₃₄ a ₁₃ , a ₂₃ , a ₁₄ , a ₂₄
ML ₆	octahedral	180.0°	a ₁₂ , a ₃₄ , a ₅₆
		90.0°	all others

^a The notation a₁₂ describes the angle between atom 1, the metal, and atom 2. The atom indices correspond to Figure 3. b_m is the mean of the four angles between the apical bond and the basal bonds in square pyramid geometries.

ligands was determined, the geometry of the metal centers was evaluated. The coordination states allowed include octahedral, Figure 3 structure 1, trigonal bipyramid, Figure 3 structure 2, square pyramid, Figure 3 structure 3, tetrahedral, Figure 3 structure 4, square planar, Figure 3 structure 5, and tetrahedral plus a nonbonded contact, Figure 3 structure 6. From Figure 3 one can see that the coordination number alone is not enough to assign a metal geometry. Thus, the root-mean-square deviation (rmsd) of the metal coordination sphere angles from those found in a regular polyhedron was calculated. Equation 4 was used to distinguish between square planar and tetrahedral geometries with the ideal angles used in Table 4. Likewise, eqs 5 and 6 were used for five- and six-coordinate metals, respectively. The atom indices in Table 4 correspond to those atoms in Figure 3. This indexing is useful to differentiate between axial/equatorial and cis/trans ligands. The coordination state with the lowest rmsd was assigned to the metal and its ligands.

Table 5. Tetrahedral Zinc Primary Ligating Residues^a

	N	bond	min	1st Q	median	mean	3rd Q	max	std dev
CCCC	3284	Zn-S	1.424	2.294	2.338	2.338	2.389	2.805	0.1218
CCCH	1041	Zn-S	1.448	2.284	2.332	2.332	2.382	3.047	0.1089
CCHH	334	Zn-S	1.908	2.234	2.295	2.301	2.361	2.795	0.1289
CHHH	14	Zn-S	2.18	2.27	2.296	2.344	2.39	2.608	0.1364
CCCH	347	Zn-N	1.833	2.056	2.124	2.132	2.2	2.525	0.1157
CCHH	334	Zn-N	1.716	2.023	2.078	2.088	2.149	2.465	0.1188
CHHH	42	Zn-N	1.778	1.964	2.034	2.056	2.113	2.486	0.1403
HHHH	12	Zn-N	1.935	2.006	2.04	2.049	2.107	2.129	0.0627
HHHO	108	Zn-O	1.359	2	2.252	2.185	2.362	2.518	0.2218
HOOO	42	Zn-O	1.866	2.092	2.268	2.233	2.384	2.543	0.1816
HOOO	78	Zn-O	1.611	2.006	2.143	2.115	2.241	2.495	0.1914
Oooo	12	Zn-O	1.781	2.004	2.158	2.135	2.3	2.428	0.1917
HHHO	324	Zn-N	1.872	2.044	2.098	2.116	2.176	2.757	0.114
HOOO	42	Zn-N	1.85	2.06	2.143	2.161	2.26	2.453	0.1455
HOOO	26	Zn-N	1.836	2.041	2.089	2.102	2.121	2.459	0.1295
HHHD	155	Zn-O	1.688	1.914	2	2.007	2.086	2.457	0.1425
HHDD	68	Zn-O	1.805	2.053	2.148	2.166	2.262	2.938	0.184
HHHD	465	Zn-N	1.604	2	2.064	2.077	2.144	2.499	0.1275
HHDD	68	Zn-N	1.959	2.101	2.184	2.192	2.302	2.46	0.128
ASP	460	Zn-O	1.688	1.958	2.044	2.077	2.165	2.988	0.1899
GLU	227	Zn-O	1.462	1.996	2.102	2.134	2.265	2.823	0.2276
HIS	825	Zn-ND	1.716	2.03	2.096	2.107	2.181	2.525	0.1256
HIS	1768	Zn-NE	1.604	2.031	2.093	2.108	2.177	2.757	0.1228

^a Bond Lengths are in Ångstroms. One-letter amino acid codes are used: C, CYS; H, HIS; O, HOH; D, ASP. N is the number of bond instances. Min and max are the minimum and maximum bond lengths, respectively. The 1st Q, 3rd Q, mean, median, and standard deviation are statistical parameters to describe the bond length distribution.

the PDB representative of each environment was chosen. FFs were built using the B3LYP DFT method with the 6-31G* basis set.⁷³ The resulting FFs were stored in xml format for later use within MTK++. For each environment a new Zn residue type was created, while new 28 amino acid residues were also added. The atom type of the bonding atom within the residue was also changed as shown in Table 6.

Two models of each representative Zn cluster were generated. For example, the two models of Zn-CCCC from PDB ID 1A5T are shown in Figure 8 (all other models are displayed in the Supporting Information).

The average optimized Zn-S bond length within Zn-CCCC was 2.43 Å, which is higher than the mean value from the survey of the PDB but within one standard deviation. The corresponding mean bond force constant is 100.7 (71.3) kcal/(mol Å²) calculated using the “traditional” and Seminario methods, respectively. The mean S-Zn-S and C-S-Zn angles are 109.1° and 101.8° with average force constants of 15.0 (73.52) and 81.5 (113.5) kcal/(mol rad²), respectively.

Structures 1A73 and 2GIV from the PDB were used as representative structures of the Zn-CCCH cluster. The delta nitrogen of His is bound to the zinc atom in 1A73, while the epsilon nitrogen is bound in 2GIV. The average Zn-S bond length was determined at 2.35 Å, which is in good agreement with the value determined from the PDB survey. The Zn-S bond lengths are shorter in Zn-CCCH than they are in the Zn-CCCC cluster, and this corresponds to the change in force constant from 100.7 (71.3) to 143.3 (105.4) kcal/(mol Å²). The mean S-Zn-S, S-Zn-N, and C-S-Zn angles for 1A73 and 2GIV clusters are 115.2°/116.4°, 102.9°/101.3°, and 102.5°/101.8° with an average force constants of 13.7 (80.0)/10.4 (72.2), 21.9 (75.7)/15.2 (70.96), and 78.6 (110.12)/65.2 (104.56) kcal/(mol rad²), respectively.

The 1A1F structure from the PDB was used as a representative structure for the Zn-CCHH cluster. The

average Zn-S and Zn-N bond lengths are 2.31 and 2.09 Å with corresponding force constants of 181.5 (143.0) and 147.1 (100.8) kcal/(mol Å²), respectively. The average value of the Zn-S bond length from the PDB is 2.301 Å, while the mean Zn-N value is 2.088 Å, which are in excellent agreement with the calculated values. Both the Zn-S and the Zn-N bonds are shorter than the previous clusters, and this is reflected by the fact that stronger force constants been observed computationally. The mean S-Zn-N, C-S-Zn, and C-N-Zn angles for the 1A1F cluster are 103.2°, 105.1°, and 126.6° with average force constants of 12.5 (78.3), 69.3 (87.7), and 34.5 (111.22) kcal/(mol rad²), respectively.

The final Zn center considered in this study which contains a cysteine residue was Zn-CHHH. The 1CK7 structure from the PDB was used to model the Zn-CHHH cluster. Two models of this cluster were built using MCPB, and the Zn-S bond length was determined as 2.26 Å with a force constant of 186.2 (168.6) kcal/(mol Å²). The mean Zn-N bond length is 2.05 Å with a force constant of 146.2 (127.8) kcal/(mol Å²). The mean N-Zn-N and S-Zn-N angles are 105.8° and 113.0° with force constants of 2.8 (75.4) and 3.3 (75.8) kcal/(mol rad²), respectively.

There are a very small number of zinc ions surrounded by four histidine residues in the PDB. However, to complete this computational study the bond and angle force constants were determined using 1PB0 as a starting geometry. The average Zn-N bond distance was determined at 2.01 Å with a force constant of 217.6 (157.0) kcal/(mol Å²). The angles of the Zn center are 109.5° with a force constant of 6.1 (70.5) kcal/(mol rad²).

It is evident there are clear trends in the calculated bond lengths and force constants described above. The bond lengths of Zn-S through the series CCCC, CCCH, CCHH, and CHHH correlate with the calculated force constants with an R² value of 0.97 (0.96) as seen in Figure 9 (top). The

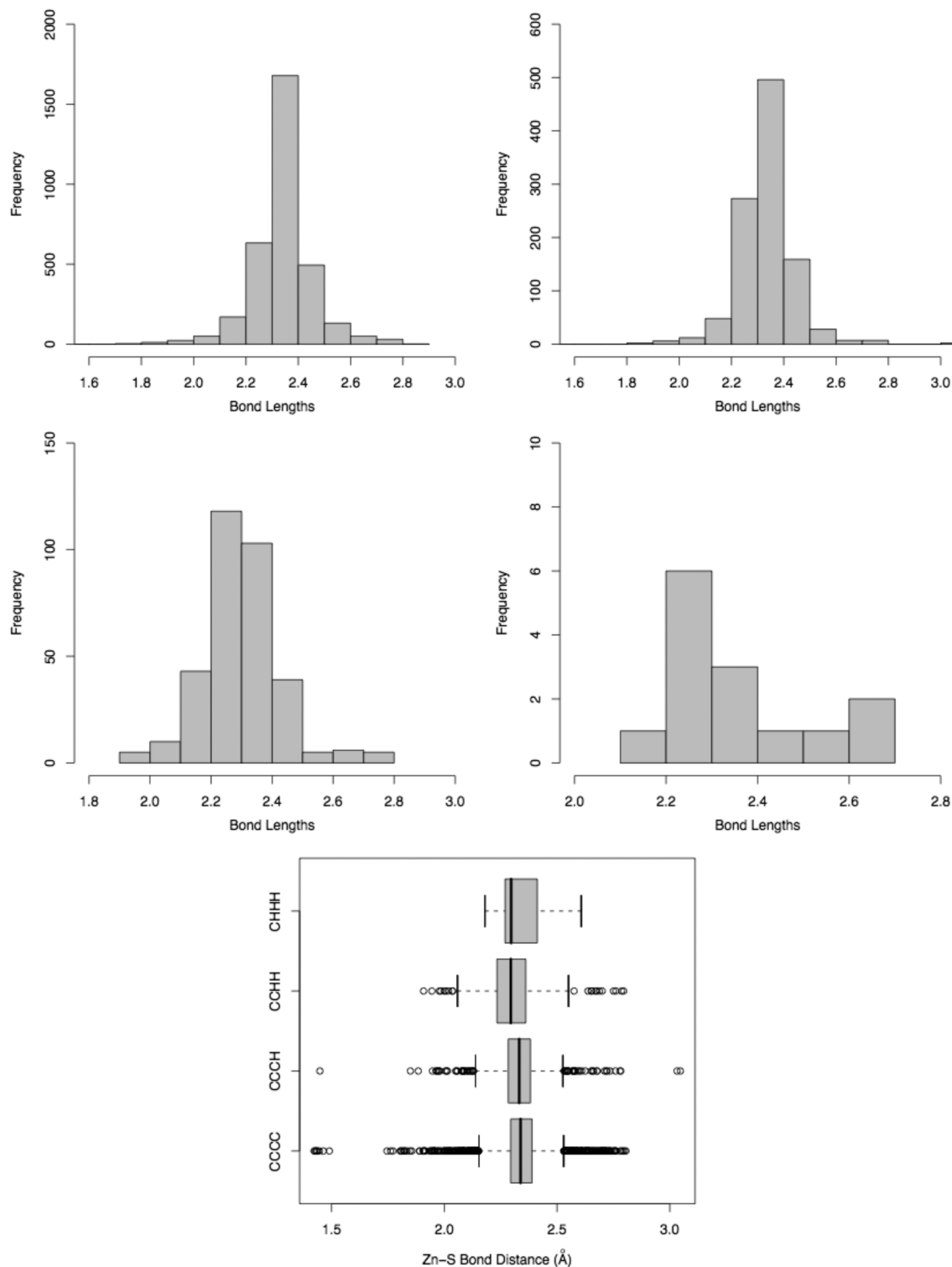


Figure 6. Zn-S bond length distributions in CCCC (top left), CCHH (top right), CCHH (middle left), and CHHH (middle right) tetrahedral environments, and a box plot summarizing all four environments (bottom).

Zn-N bond lengths and force constants correlate with an R^2 of 0.95 (0.98) as shown in Figure 9 (bottom). It is worth noting here that Zn donor bond lengths differ within the various environments. Thus, having a single Zn-S or Zn-N bond equilibrium and force constant value will not work appropriately. The proposed solution to this problem is to store all Zn bond types and assign the parameters in an automatic manner within the metal center perception algorithm of MTK++.

The average angle size and force constants of S-Zn-S are smaller and stronger, $109.5^\circ/15.0$ (73.5) kcal/(mol rad²) in the CCCC cluster compared to those of the CCHH cluster where values of $135.0^\circ/9.7$ (104.4) kcal/(mol rad²) were determined. The N-Zn-N angles of the CCHH, CHHH, and HHHH clusters lie between 105.8° and 109.5° with force constants between 2.8 (75.4) and 8.4 (58.0) kcal/(mol rad²). The experimental force constant of N-Zn-N was reported to be approximately 5.0 kcal/(mol rad²), which is in good

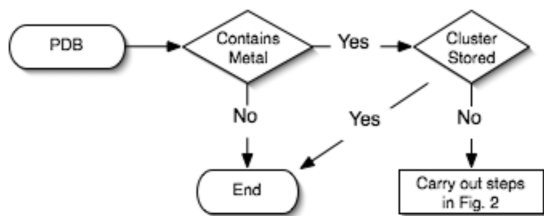


Figure 7. ZAFF flow diagram. This illustration demonstrates when a metalloprotein structure is downloaded from the PDB and an equivalent metal site is stored the MTK++ package has the ability to assign parameters to carry out MD simulations.

Table 6. List of PDB codes, Zn Environments, Residue, and Atom Type Names That Were Used To Create ZAFF

PDB	environment	residue/type			
1A5T	Zn-CCCC	ZN1	CY1/S1		
1A73	Zn-CCCH	ZN2	CY2/S2	HE1/N1	
2GIV	Zn-CCCH	ZN3	CY3/S3	HD1/N2	
1A1F	Zn-CCHH	ZN4	CY4/S4	HD2/N3	
1CK7	Zn-CHHH	ZN5	CY5/S5	HD3/N4	
1CA2	Zn-HHHO	ZN6	HD4/N5	HD5/N6	HE2/N7 WT1/O1
1CA2	Zn-HHHO	ZN7	HD6/N8	HD7/N9	HE3/N0 OH1/O2
1U0A	Zn-HHDD	ZN8	HE4/E1	AP1/D1	
2USN	Zn-HHHD	ZN9	HD8/NQ	HD9/NP	HE5/E2 AP2/D2
1PB0	Zn-HHHH	Z10	HDD/NR		
1VLI	Zn-HHOO	Z11	HDA/NT	WT2/O3	
1L3F	Zn-HOOO	Z12	HE6/NY	WT3/O4	

agreement with those calculated by the “traditional” method.³ It has been reported that this angle force constant is too weak to prevent the angle opening beyond the ideal tetrahedral angle in MD simulations, and in the past arbitrary scaling factors have been applied to prevent this from occurring.^{3,41} However, this does not seem to be the case for the Seminario method, and so the need to arbitrarily scale values is negated.

The variation of bond distances and angles of zinc clusters containing histidine residues and water molecules were determined. The 1CA2 structure was used to represent the Zn-HHHO cluster which is a structure of human Carbonic Anhydrase II (HCA II). HCA II is a catalytic center for the conversion of CO₂ into bicarbonate. Therefore, to account for both the water and the hydroxyl states two FFs were evaluated. It is of no surprise that the bond lengths and associated force constants of the two systems are different. The Zn-O bond is longer in the case of water binding, while the Zn-N bonds are shorter due to the strength of the hydroxyl bond. The accompanying force constants are also considerably different. The Zn-O bond force constants changes from 120.3 (82.86) to 394.7 (303.6) kcal/(mol Å²) upon removal of a proton, while the Zn-N force constant becomes weaker from 248.4 (184.4) to 194.4 (131.3) kcal/(mol Å²). These calculated equilibrium bond lengths and force constants are different from those published by Hoops et al.; however, the QM methods used to generate the numbers also differ. The Zn-O bond lengths of the HHHO clusters in the PDB have a large standard deviation of 0.22 Å with a mean value of 2.19 Å, confirming that both states exist. The calculated angles and force constants for this cluster are in good agreement with those published previously, except for the H-O-Zn angle force constant that was

scaled to a higher value to ensure trajectory stability during MD simulation.

The 1VLI structure from the PDB was used to investigate the strength of bond and angle force constants of the Zn-HHOO cluster. Again, the MCPB program was used to build the models required for parametrization. The equilibrium bond length of Zn-O was calculated as 1.95 Å, which is approximately 0.4 Å shorter than the bond length for the Zn-HHOO cluster. This contradicts the trend from the PDB survey. Plausible reasons for this discrepancy include the small number of data points for the Zn-HHOO cluster in the PDB and the large standard deviation value of 0.15 Å. The angle force constants calculated for this cluster are of similar magnitude to those calculated for the Zn-HHOO cluster.

The final tetrahedral environment containing histidine residues and water molecules was the HOOO cluster from PDB ID 1L3F. The average Zn-O and Zn-N bond lengths are 2.01 and 1.93 Å, respectively, which are shorter distances than those in the Zn-HHHO and Zn-HHOO clusters, agreeing with the experimental means from the PDB.

Two clusters containing histidine and aspartate residues were considered in this study. The 2USN and 1U0A were chosen as characteristic structures of the Zn-HHHD and Zn-HHDD environments. The PDB survey showed that the bond lengths of Zn-O bonds in H/D systems changed from 2.007 to 2.166 Å going from HHHD to HHDD, and this trend is also seen in the calculated values of these clusters.

Molecular Dynamics Validation. The partial charges of the zinc clusters were determined applying four different methods using the larger models as described above. It was difficult to determine, a priori, which method would outperform the others. Thus, MD simulations and normal-mode analysis were used to determine which combination of force constant approach (traditional and Seminario) and partial charge model (ChgModA, -B, -C, -D) was most appropriate. With this in mind, we carried out MD simulations on two different systems, one of which was a zinc finger complex (PDB ID 1A5T), and the other was a dizinc system (PDB ID 1AMP) as shown in Figure 10. The zinc finger system was used to compare the difference between different charge models, while the dizinc system was chosen to further evaluate the feasibility of our approach on a dimetal cluster.

In each of the eight simulations of 1A5T, the system was solvated in a rectangular periodic box with each side of the box having a distance of at least 8.0 Å from the closest atom from the enzyme. The TIP3P water model was used, and sodium ions were added to neutralize the system. The AMBER FF ff99SB was used to model the enzyme, while the zinc complex was modeled with the parameters generated using the MCPB package. The particle mesh ewald (PME) method was applied to handle the long-range electrostatic interactions and the default setting of an 8.0 Å cutoff for the real-space nonbond interactions and 1.0 Å grid spacing for the reciprocal space was used. All bonds with hydrogen atoms were constrained using SHAKE. The whole system was first fully minimized with a weak positional restraint in order to eliminate any close contacts. Then the temperature

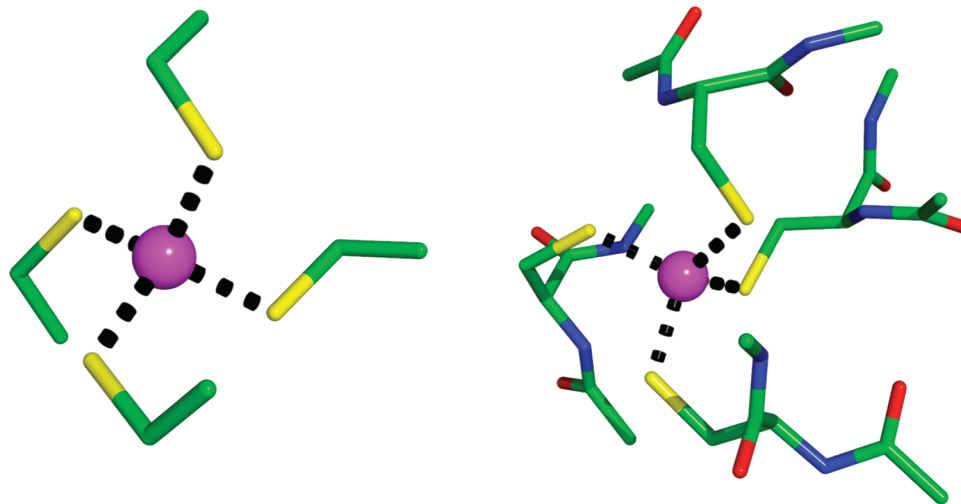


Figure 8. Zn–CCCC cluster models (PDB ID 1A5T).

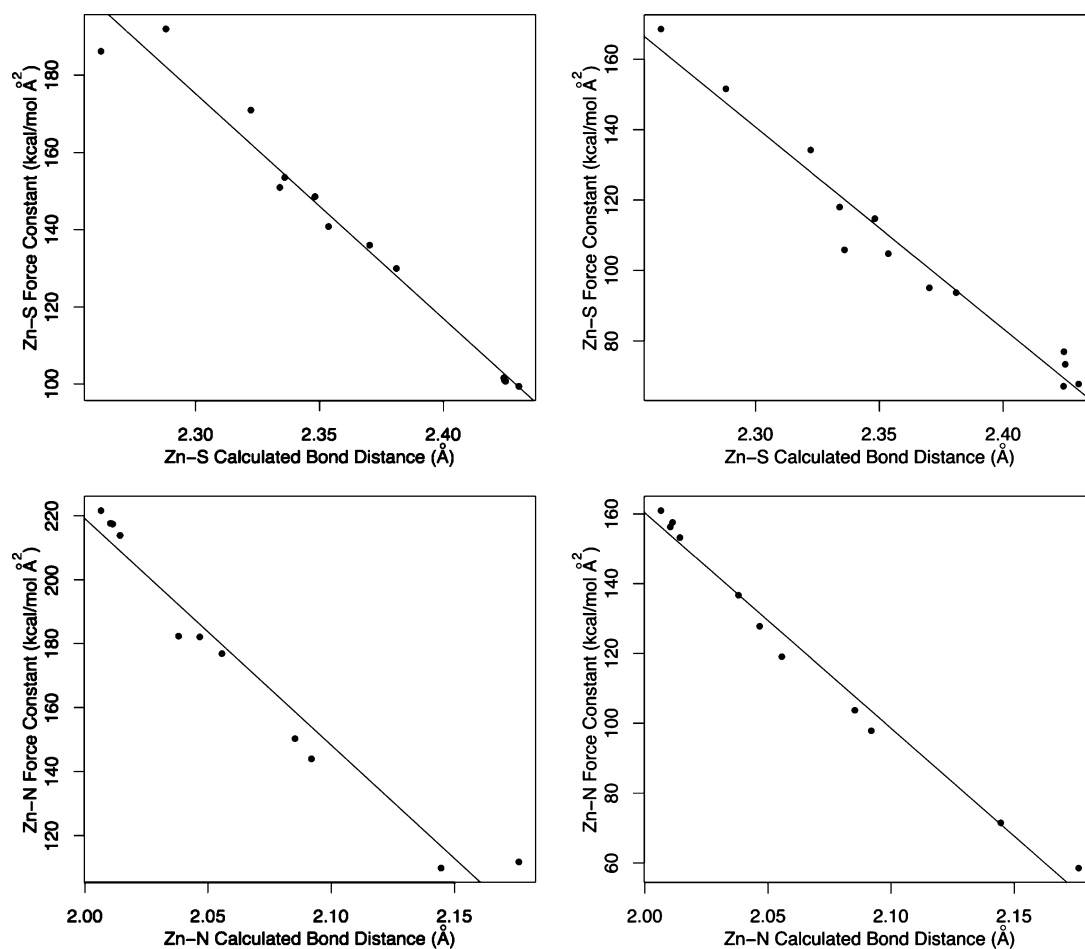


Figure 9. Correlation between (top) Zn–Cys@S and (bottom) Zn–His@N bond lengths and calculated force constants through the series CCCC, CCCH, CCHH, CHHH, and HHHH using the “traditional” and Seminario methods.

was slowly increased to 300 K during a 100 ps, 0.5 fs time step canonical ensemble simulation with the restraint force being gradually removed. This was followed by 3 ns NPT production runs at 1 atm with a 2 fs time step.

The results from eight sets of simulations with different sets of parameters were collected and carefully compared by four criteria (Table 7): (1) Average $\delta_{\text{tet/sqp}}$ by using eq 4, (2) average mean Zn–S bond lengths, (3) average rmsd of zinc complex, and (4) average rmsd of the backbone of the

enzyme. In general, the simulation which used the Seminario force constants outperformed those from the “traditional” method mainly due to the fact that the force constants of zinc-containing angles are smaller in the latter approach. The Seminario/ChgModB combination gave the most impressive performance, while Seminario/ChgModD also showed promising results. This is in keeping with the fact that the backbone atoms were used in the fitting of the torsional parameters of the FF.

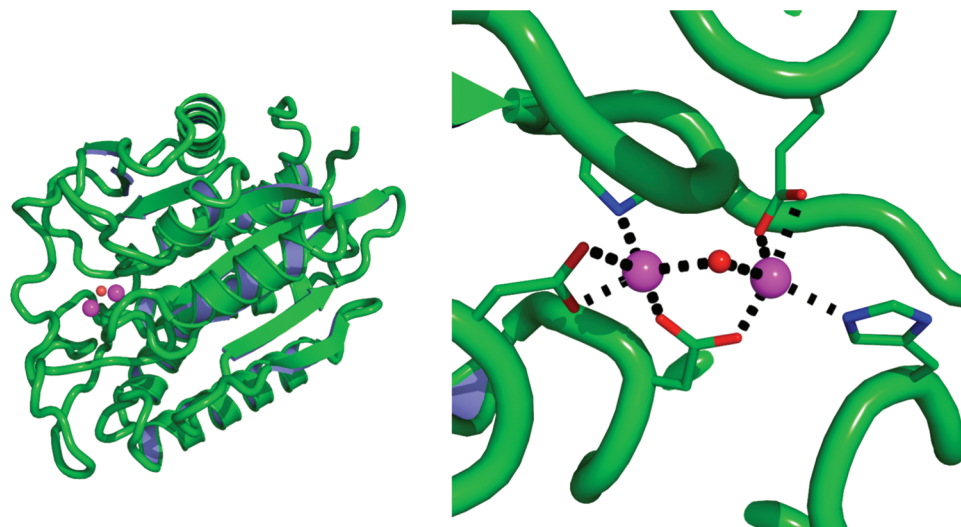


Figure 10. Cartoon representation (left) of the dizinc protein (PDB ID 1AMP) and its metal sites (right).

Table 7. Comparison of the Performance from MD Simulations Using Eight Sets of Parameters

	Seminario	traditional
	average $\delta_{\text{tet/sqp}}$	
ChgModA	4.64	8.45
ChgModB	4.37	9.27
ChgModC	4.59	9.06
ChgModD	4.42	8.33
	average mean Zn–S bonds lengths (Å)	
ChgModA	2.33	2.35
ChgModB	2.40	2.41
ChgModC	2.41	2.42
ChgModD	2.43	2.43
	average rmsd of zinc complex (Å)	
ChgModA	4.62	8.47
ChgModB	4.35	9.29
ChgModC	4.58	9.08
ChgModD	4.41	8.34
	backbone rmsd (Å)	
ChgModA	1.63	1.96
ChgModB	1.44	1.64
ChgModC	1.68	1.53
ChgModD	1.31	2.04

We also carried out normal-mode calculations on the second model of 1A5T, as shown in Figure 8, with different sets of parameters to check if the MCPB-built parameters reproduce B3LYP/6-31G*-computed results. Figure 11 shows the fitting results of frequencies calculated from normal-mode calculation with the Seminario/ChgModB set of parameters to those from QM calculation. The matching between MM normal modes and the QM ones is good despite discrepancies above 3200 cm^{-1} , which corresponds to X–H bond (X = heavy atom) stretching.

Despite the fact that around 46% of zinc structures are found to be tetrahedral, we felt it was necessary to further validate our program and strategy on a less common zinc complex, for instance, a dizinc system. As a prototypical member of the dizinc enzyme family, *Aeromonas proteolytica* aminopeptidase (PDB ID 1AMP) featuring a dizinc binding pocket offers a good choice. The same procedure as applied on 1A5T (see above) was followed, with the only difference being that only Seminario/ChgModB set of

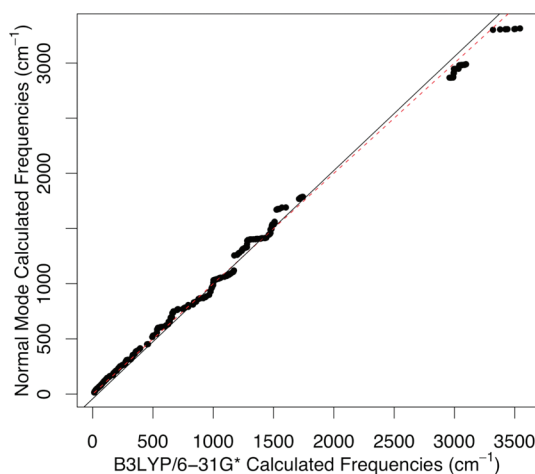


Figure 11. Correlation from 1A5T normal-mode calculation with parameters built from the Seminario model to the B3LYP/6-31G*-computed frequencies ($F^2 = 0.99$).

parameters, which gave the best performance in the 1A5T case, were built. Results from both MD simulations and normal-mode calculations were collected and examined. From Figure 12, the rmsd of the backbone and the zinc active site complex (residues HID97, ASP117, GLU152, ASP179, HID256, OH294, ZN292, ZN293) are stable throughout the simulation. Moreover, the MM normal-mode results showed a very good agreement with the QM results except in the $>3200\text{ cm}^{-1}$ region, as shown in Figure 13. Thus, our conclusions from the 1A5T test are confirmed in that MCPB provides a reliable parameter set for, at least, zinc-containing metal systems.

Conclusions

This research describes the design, development, and implementation of two programs called *pdbSearcher* and *MCPB*. The former carries out metalloprotein data mining of the Protein Data Bank. Results focused on zinc metalloproteins as a large number of proteins contain this element. We present a concept of a “cluster parameterization” where a metal ion and its coordination environment are treated as a

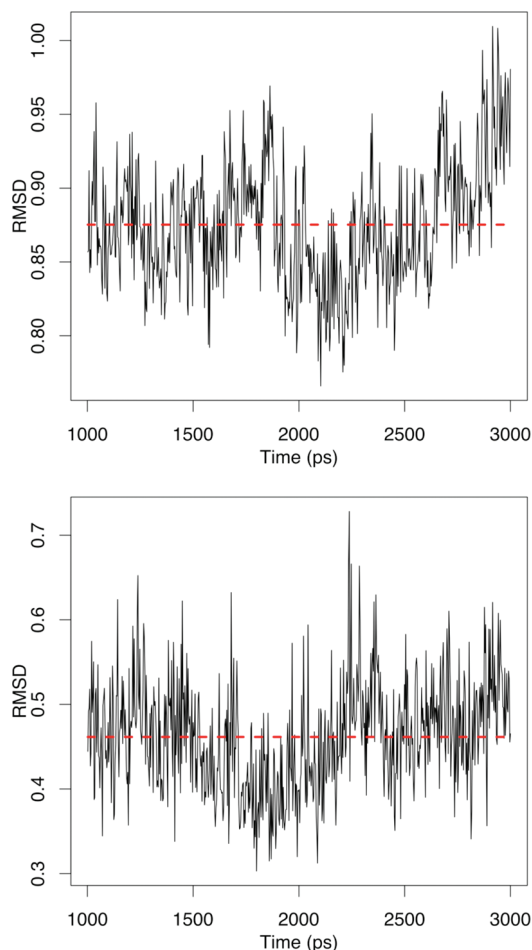


Figure 12. rmsd (in Å) of the 1AMP backbone heavy atoms (above), and rmsd of the zinc complex (bottom) over the final 2 ns.

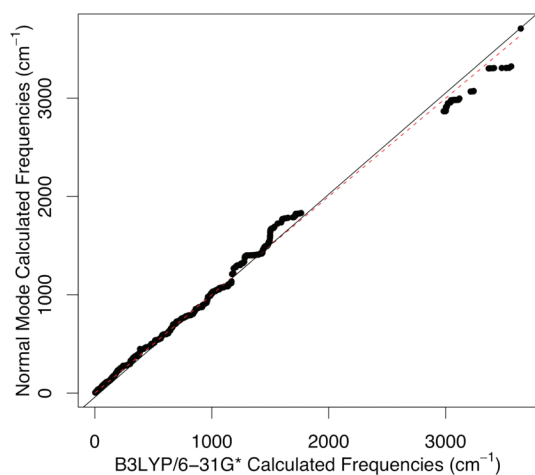


Figure 13. Fitting results from 1AMP normal-mode calculation with parameters built from the Seminario model coupled with restraint scheme 1 to B3LYP/6-31G*-computed frequencies ($R^2 = 0.99$).

single unit. The majority of Zn metalloproteins are tetrahedrally coordinated to histidine, cysteine, aspartate, glutamate residues, or water molecules. Thus, cluster models of each of these environments have been built. The distribution of bond lengths between Zn and the donor atoms of these residues was investigated, with some short Zn–S bonds

highlighted which may be due to errors during crystal structure refinement. Not unexpected periodicities of Zn–S and Zn–N bonds were observed (bond lengths change as a function of coordination environment), highlighting the need for a detailed analysis of a metallocluster behavior before assigning force constants and equilibrium bond lengths. The notion of a standard Zn–N or Zn–S force constant, for example, within the AMBER FF is not supported by this work.

The MCPB program was used to build, prototype, and validate AMBER-like force fields using the bonded plus electrostatics model for metalloproteins that can be added to the AMBER suite of programs. MCPB was used to investigate the environmental effects on bond lengths, angles, and bond and angle force constants using 10 unique metal coordination environments. These included Zn bound to CCCC, CCCH, CCHH, CHHH, HHHH, HHHO, HHOO, HOOO, HHHD, and HHDD clusters. A zinc AMBER force field (ZAFF) library was created to store these FF parameters in a convenient way as to allow later use with different metalloproteins than those used in the parametrization.

This work has many current and future uses. The equilibrium bond lengths and angles can be used to aid the refinement of Zn metalloprotein X-ray crystal structures. The models developed can be used by nonexperts in studies of zinc metalloproteins for the coordination environments examined. The present work also provides the foundation to develop new approaches to the accurate model of metalloproteins using nondiagonal force fields along with advanced electronic modeling. The MCPB program allows for rapid development, limited by the cost of the ab initio or DFT calculations, of FF parameters for metalloproteins which could have many uses in drug design projects, for example, where the target structure contains a metal ion; the feasibility of building parameters with this program has been tested via extensive MD simulations. This program also provides a platform where nonexpert users can develop metalloprotein FF parameters which until now was not available. We are currently using the approaches outlined herein to explore next-generation force fields for a number of metalloprotein systems.

Acknowledgment. We thank the NIH (GM044974 and GM066859) for financial support of this research.

Supporting Information Available: The equations used to determine the Seminario force constants are outlined; graphical representation of the Zn cluster models used in this study; Zn–N and Zn–O bond distributions and box plots; ZAFF xml parameter files (Seminario/ChgModB) that can be used with MTK++/AMBER. This material is available free of charge via the Internet at <http://pubs.acs.org>. Additionally, the source code of MTK++, pdbSearcher, and MCPB are available from the authors upon request at <http://www.qtp.ufl.edu/~kmmprogs/>.

References

- (1) Hancock, R. D. *Acc. Chem. Res.* **1990**, *23*, 253.
- (2) Hancock, R. D. *Prog. Inorg. Chem.* **1989**, *37*, 187.

- (3) Hoops, S. C.; Anderson, K. W.; Merz, K. M., Jr. *J. Am. Chem. Soc.* **1991**, *113*, 8262.
- (4) Bayly, C. I.; Cieplak, P.; Cornell, W. D.; Kollman, P. A. *J. Phys. Chem.* **1993**, *97*, 10269.
- (5) Li, J. B.; Zhu, T. H.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. A* **1998**, *102*, 1820.
- (6) Stote, R. H.; Karplus, M. *Proteins* **1995**, *23*, 12.
- (7) Sakharov, D. V.; Lim, C. *J. Am. Chem. Soc.* **2005**, *127*, 4921.
- (8) Aqvist, J.; Warshel, A. *J. Mol. Biol.* **1992**, *224*, 7.
- (9) Pang, Y. P. *Proteins* **2001**, *45*, 183.
- (10) Pang, Y. P.; Xu, K.; Yazal, J. E.; Prendergas, F. G. *Protein Sci.* **2000**, *9*, 1857.
- (11) Vedani, A.; Huhta, D. W. *J. Am. Chem. Soc.* **1990**, *112*, 4759.
- (12) Vedani, A.; Huhta, D. W.; Jacober, S. P. *J. Am. Chem. Soc.* **1989**, *111*, 4075.
- (13) Gresh, N. *Curr. Pharm. Des.* **2006**, *12*, 2121.
- (14) Gresh, N.; Piquemal, J. P.; Krauss, M. *J. Comput. Chem.* **2005**, *26*, 1113.
- (15) Rappe, A. K.; Casewit, C. J.; Colwell, K. S.; Goddard, W. A.; Skiff, W. M. *J. Am. Chem. Soc.* **1992**, *114*, 10024.
- (16) Rappe, A. K.; Colwell, K. S.; Casewit, C. J. *Inorg. Chem.* **1993**, *32*, 3438.
- (17) Sirovatka, J. M.; Rappe, A. K.; Finke, R. G. *Inorg. Chim. Acta* **2000**, *300*, 545.
- (18) Brandt, P.; Norrby, T.; Akermark, E.; Norrby, P. O. *Inorg. Chem.* **1998**, *37*, 4120.
- (19) Marques, H. M.; Brown, K. L. *THEOCHEM* **1995**, *340*, 97.
- (20) Brown, K. L.; Zou, X.; Marques, H. M. *J. Mol. Struct.: THEOCHEM* **1998**, *453*, 209.
- (21) Marques, H. M.; Brown, K. L. *Coord. Chem. Rev.* **1999**, *192*, 127.
- (22) Marques, H. M.; Ngoma, B.; Egan, T. J.; Brown, K. L. *J. Mol. Struct.* **2001**, *561*, 71.
- (23) Aqvist, J.; Warshel, A. *J. Am. Chem. Soc.* **1990**, *112*, 2860.
- (24) Ryde, U. *Protein Sci.* **1995**, *4*, 1124.
- (25) Ryde, U. *Proteins: Struct., Funct., Genet.* **1995**, *21*, 40.
- (26) Hancock, R. D.; Weaving, J. S.; Marques, H. M. *J. Chem. Soc., Chem. Commun.* **1989**, 1176.
- (27) Marques, H. M.; Brown, K. L. *Coord. Chem. Rev.* **2002**, *225*, 123.
- (28) Marques, H. M.; Cukrowski, I. *Phys. Chem. Chem. Phys.* **2002**, *4*, 5878.
- (29) Skopec, C. E.; Robinson, J. M.; Cukrowski, I.; Marques, H. M. *J. Mol. Struct.* **2005**, *738*, 67.
- (30) Skopec, C. E.; Cukrowski, I.; Marques, H. M. *J. Mol. Struct.* **2006**, *783*, 21.
- (31) Norrby, P. O.; Liljefors, T. *J. Comput. Chem.* **1998**, *19*, 1146.
- (32) Norrby, P. O.; Brandt, P. *Coord. Chem. Rev.* **2001**, *212*, 79.
- (33) Allinger, N. L.; Yuh, Y. H.; Lii, J. H. *J. Am. Chem. Soc.* **1989**, *111*, 8551.
- (34) Sindhikara, D. J.; Roitberg, A. E.; Merz, K. M. *Biochemistry* **2009**, *48*, 12024.
- (35) Lin, F.; Wang, R. *J. Chem. Theory Comput.* **2010**, *6*, 1852.
- (36) Merz, K. M., Jr. *J. Am. Chem. Soc.* **1991**, *113*, 406.
- (37) Merz, K. M., Jr.; Murcko, M. A.; Kollman, P. A. *J. Am. Chem. Soc.* **1991**, *113*, 4484.
- (38) Diaz, N.; Suarez, D.; Merz, K. M., Jr. *J. Am. Chem. Soc.* **2001**, *123*, 9867.
- (39) Suarez, D.; Brothers, E. N.; Merz, K. M., Jr. *Biochemistry* **2002**, *41*, 6615.
- (40) Suarez, D.; Diaz, N.; Merz, K. M., Jr. *J. Comput. Chem.* **2002**, *23*, 1587.
- (41) Cui, G. L.; Wang, B.; Merz, K. M., Jr. *Biochemistry* **2005**, *44*, 16513.
- (42) Ullmann, G. M.; Knapp, E. W.; Kostic, N. M. *J. Am. Chem. Soc.* **1997**, *119*, 42.
- (43) De Kerpel, J. O. A.; Ryde, U. *Proteins: Struct., Funct., Genet.* **1999**, *36*, 157.
- (44) Olsson, M. H. M.; Ryde, U. *J. Biol. Inorg. Chem.* **1999**, *4*, 654.
- (45) Remenyi, R.; Comba, P. *J. Inorg. Biochem.* **2001**, *86*, 397.
- (46) Comba, P.; Remenyi, R. *J. Comput. Chem.* **2002**, *23*, 697.
- (47) Estiu, G.; Merz, K. M., Jr. *Biochemistry* **2006**, *45*, 4429.
- (48) Estiu, G.; Suarez, D.; Merz, K. M., Jr. *J. Comput. Chem.* **2006**, *27*, 1240.
- (49) Collins, J. R.; Camper, D. L.; Loew, G. H. *J. Am. Chem. Soc.* **1991**, *113*, 2736.
- (50) Collins, J. R.; Du, P.; Loew, G. H. *Biochemistry* **1992**, *31*, 11166.
- (51) Yao, S. J.; Plastaras, J. P.; Marzilli, L. G. *Inorg. Chem.* **1994**, *33*, 6061.
- (52) Branco, R. J. F.; Fernandes, P. A.; Ramos, M. J. *J. Phys. Chem. B* **2006**, *110*, 16754.
- (53) Case, D. A.; Darden, T. A.; Cheatham, III, T. E.; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Crowley, M.; Walker, R. C.; Zhang, W.; Merz, K. M.; Wang, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Kolossváry, I.; Wong, K. F.; Paesani, F.; Vanicek, J.; Wu, X.; Brozell, S. R.; Steinbrecher, T.; Gohlke, H.; Yang, L.; Tan, C.; Mongan, J.; Hornak, V.; Cui, G.; Matthews, D. H.; Seetin, M. G.; Sagui, C.; Babin, V.; Kollman, P. A. (2008), Amber 10, University of California, San Francisco.
- (54) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz Jr., K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. *J. Am. Chem. Soc.* **1995**, *117*, 5179.
- (55) Brooks, B. R.; Brucoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* **1983**, *4*, 187.
- (56) Harding, M. M. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **1999**, *55*, 1432.
- (57) Harding, M. M. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2000**, *56*, 857.
- (58) Harding, M. M. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2001**, *57*, 401.
- (59) Harding, M. M. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2002**, *58*, 872.
- (60) Harding, M. M. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2004**, *60*, 849.
- (61) Harding, M. M. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2006**, *62*, 678.

- (62) Bondi, A. *J. Phys. Chem.* **1964**, *68*, 441.
- (63) Aqvist, J. *J. Phys. Chem.* **1990**, *94*, 8021.
- (64) Babu, C. S.; Lim, C. *Chem. Phys. Lett.* **1999**, *310*, 225.
- (65) Babu, C. S.; Lim, C. *J. Phys. Chem. B* **1999**, *103*, 7958.
- (66) Babu, C. S.; Lim, C. *J. Phys. Chem. A* **2006**, *110*, 691.
- (67) Vaiana, A. C.; Schulz, A.; Wolfrum, J.; Sauer, M.; Smith, J. C. *J. Comput. Chem.* **2003**, *24*, 632.
- (68) 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.; 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.
- (69) Seminario, J. M. *Int. J. Quantum Chem.* **1996**, *60*, 1271.
- (70) Singh, U. C.; Kollman, P. A. *J. Comput. Chem.* **1984**, *5*, 129.
- (71) Besler, B. H.; Merz, K. M., Jr.; Kollman, P. A. *J. Comput. Chem.* **1990**, *11*, 431.
- (72) Cieplak, P.; Cornell, W. D.; Bayly, C.; Kollman, P. A. *J. Comput. Chem.* **1995**, *16*, 1357.
- (73) Siegbahn, P. E. M.; Borowski, T. *Acc. Chem. Res.* **2006**, *39*, 729.

CT1002626

Structural Fluctuations in Enzyme-Catalyzed Reactions: Determinants of Reactivity in Fatty Acid Amide Hydrolase from Multivariate Statistical Analysis of Quantum Mechanics/Molecular Mechanics Paths

Alessio Lodola,^{*,†} Jitnapa Sirirak,[‡] Natalie Fey,[‡] Silvia Rivara,[†] Marco Mor,[†] and Adrian J. Mulholland^{*,‡}

Dipartimento Farmaceutico, Università degli Studi di Parma, 43124 Parma, Italy, and Centre for Computational Chemistry, School of Chemistry, University of Bristol, Bristol BS8 1TS, United Kingdom

Received May 19, 2010

Abstract: The effects of structural fluctuations, due to protein dynamics, on enzyme activity are at the heart of current debates on enzyme catalysis. There is evidence that fatty acid amide hydrolase (FAAH) is an enzyme for which reaction proceeds via a high-energy, reactive conformation, distinct from the predominant enzyme–substrate complex (Lodola et al. *Biophys. J.* 2007, 92, L20–22). Identifying the structural causes of differences in reactivity between conformations in such complex systems is not trivial. Here, we show that multivariate analysis of key structural parameters can identify structural determinants of barrier height by analysis of multiple reaction paths. We apply a well-tested quantum mechanics/molecular mechanics (QM/MM) method to the first step of the acylation reaction between FAAH and oleamide substrate for 36 different starting structures. Geometrical parameters (consisting of the key bond distances that change during the reaction) were collected and used for principal component analysis (PCA), partial least-squares (PLS) regression analysis, and multiple linear regression (MLR) analysis. PCA indicates that different “families” of enzyme–substrate conformations arise from QM/MM molecular dynamics simulation and that rarely sampled, catalytically significant conformational states can be identified. PLS and MLR analyses allowed the construction of linear regression models, correlating the calculated activation barriers with simple geometrical descriptors. These analyses reveal the presence of two fully independent geometrical effects, explaining 78% of the variation in the activation barrier, which are directly correlated with transition-state stabilization (playing a major role in catalysis) and substrate binding. These results highlight the power of statistical approaches of this type in identifying crucial structural features that contribute to enzyme reactivity.

Introduction

Enzymes exhibit a wide range of internal motions, some of which are essential for their activity.^{1,2} The relationship among these conformational changes, structural fluctuations, and

enzyme catalysis is a current central issue in enzymology.^{3–5} Debates on the possible role of dynamics in enzyme catalysis are very heated. Different authors have suggested different definitions of the term “dynamics”. A rigorous definition is that dynamical effects are any corrections to transition-state theory predictions, due to, e.g., recrossing of the barrier: the indications are that such effects on the rates of enzyme-catalyzed reactions are relatively small.⁶ There is compelling evidence that transition-state theory provides a good basis

* Corresponding author e-mail: alessio.lodola@unipr.it (A.L.); adrian.mulholland@bristol.ac.uk (A.J.M.).

[†] Università degli Studi di Parma.

[‡] University of Bristol.

for understanding chemical reactions in enzymes^{7–9} and other catalysts.¹⁰ Other authors identify the effects of dynamics with structural fluctuations that do not “drive” reaction, but may be crucial in permitting efficient reaction. Such fluctuations are subsumed into the free energy of activation in terms of transition-state theory. Conformational changes are an essential part of many enzyme catalytic cycles,¹¹ allowing, for example, efficient product release or preventing loss or hydrolysis of intermediates. The ability of many enzymes to undergo specific structural changes rapidly appears to have been evolved as an essential part of their overall catalytic efficiency. Subtle changes in the enzyme active site geometry can be crucial for efficient reaction.¹² They may also have a functional role, for example in enhancing substrate specificity. In some cases, local fluctuations of the enzyme active site could couple to substrate fluctuations, significantly reducing the barrier heights.¹³ In other examples, fluctuations between different conformations of the enzyme–substrate complex could simply reflect a relatively flat free energy surface along a specific reaction coordinate.¹⁴ However, it is also possible that a distinct high-energy (lowly populated) conformation of the enzyme–substrate complex may dominate its reactivity. Fatty acid amide hydrolase (FAAH) appears to be such a case.¹⁵

Detailed insight into the functioning of enzymes can be obtained from molecular simulations.^{16,17} Simulations of enzyme–substrate complexes (e.g., by applying molecular mechanics (MM) force fields) can provide insight into some potentially important determinants of reactivity, such as the proximity of reacting groups or the conformational behavior of substrates.^{11,18} However, for a detailed understanding of a catalytic process, both the identification of transition-state (TS) structures and their energies relative to the reactants are required. A well-established approach is hybrid quantum mechanics/molecular mechanics (QM/MM) methods,^{19,20} which combine the simplicity and speed of the MM treatment of the protein structure with the flexibility and power of a quantum chemical treatment, thus allowing computational modeling of bond breaking and making in enzymes.²¹

In the QM/MM approach, a small region of the active site (where the reaction happens) is treated by a QM electronic structure method. This interacts with the protein and solvent environment, which are described by an empirical MM force field.²² Most QM/MM studies of enzyme-catalyzed reactions²³ make use of a protein–substrate complex, built starting from the crystallographic coordinates of a protein–ligand structure (e.g., from the Protein Data Bank (PDB)). A refinement procedure, including addition of hydrogen atoms and missing side chains, solvation (by adding water molecules), and thermal equilibration of the resulting complex by molecular dynamics (MD) simulations, is usually carried out before the catalytic process is modeled.²⁴ Once the structure has been prepared, a QM/MM potential, in combination with a conformational sampling algorithm, can be applied to identify a TS for the reaction and to estimate its activation free energy.²⁵ However, this approach can be computationally demanding, particularly when a biochemical process is complex (e.g., composed of several chemical steps)

and/or multiple possible mechanisms need to be explored. In some of these cases, a suitable strategy to investigate enzyme-catalyzed reactions is the “adiabatic mapping” approach,²⁶ where the potential energy surface (PES) of a reaction is calculated by a simple energy minimization along an approximate reaction coordinate.²⁷ With this approach, only a single starting structure is employed to explore the PES, with the assumption that this structure is a reasonable representation of the conformational space of the reacting enzyme–substrate complex. Adiabatic mapping has been useful in elucidating several enzyme mechanisms,²⁸ but is not always appropriate (e.g., for reactions involving large changes of solvation or movements of charge).^{29,30} In enzymes for which adiabatic mapping is useful (i.e., those reactions for which a potential energy surface is similar to the free energy surface), such calculations must be performed and analyzed carefully: for example, the selection of an erroneous starting structure can lead to a structurally or chemically incorrect mechanism being modeled.³¹ Geometrical fluctuations of the active site can significantly affect the overall energetic barrier, suggesting that the use of different protein conformations as starting geometries is important.³² As emphasized in recent literature,^{23,31} in QM/MM potential energy surface calculations several “representative” transition-state structures should be considered along with their corresponding minima. The resulting multiple PESs allow detailed analysis of the structural features that affect the barrier height of an enzyme-catalyzed reaction. While this approach does not give complete configurational sampling, it will at least partly capture the influence of the conformational diversity of the environment on the reaction. An expedient way to generate a selection of configurations is to take snapshots from an MD trajectory and use them as starting structures in subsequent QM/MM optimizations.^{23,32}

FAAH is an important enzyme in the central nervous system, responsible for the (deactivating) hydrolysis of the endocannabinoids and other bioactive lipid amides.³³ We have previously identified a likely mechanism for the first step of the acylation reaction of FAAH with oleamide (OA) using QM/MM methods employing a limited number (four) of starting structures.³⁴ This mechanism is also supported by QM/MM Monte Carlo simulations³⁵ and kinetic experiments.³⁶ Here, we examine in detail conformational effects in FAAH, building on our multiple QM/MM reaction path approach, by extracting 36 snapshots from an MD trajectory as starting points for modeling the FAAH–OA reaction at the PM3-CHARMM22 level.^{37,38} Although semiempirical methods are known to overestimate many energy barriers,³⁹ including FAAH,³⁴ calculations at this level of theory have been shown to be useful in many cases both for comparing different reaction mechanisms⁴⁰ and for modeling multiple reaction pathways with several substrates.⁴¹ Here, the focus is on a comparison of barriers for the same mechanism, and such relative barriers will be considerably more accurate.¹⁵ Thus, the present approach should allow a detailed exploration of the effect of structural fluctuations on the calculated energy barrier. We focused our investigation on the key points of the PES (substrate complex and TS) and analyzed the variation of geometrical parameters (consisting of the

key distances and angles that change during the course of the reaction) for all starting geometries and TS structures. To extract meaningful information from QM/MM calculations, we have employed a range of statistical approaches: (a) principal component analysis (PCA),⁴² (b) partial least-squares (PLS) regression analysis,^{43,44} (c) multiple linear regression (MLR) analysis.⁴⁵ These are widely used statistical techniques⁴⁶ in quantitative structure–activity relationships (QSARs).^{47–49} Although PCA and/or PLS has been applied to the field of QM calculations on small model systems,^{50–52} to the best of our knowledge this is the first time where several multivariate statistical techniques have been employed to relate the changes in energy to distances and angles in QM/MM calculations of an enzyme. The analyses revealed the presence of two fully independent geometrical effects directly correlated with substrate binding and TS stabilization controlling the rate of the reaction. These results highlight the power of statistical approaches in identifying crucial features contributing to enzyme catalysis.

Methods: Multivariate Data Analysis

PCA, PLS, and MLR statistical models were generated to ascertain how the structural parameters of the key species in the reaction change from snapshot to snapshot and how these changes affect the energy of the system. PCA is a statistical method for reducing the amount of data to be analyzed by exploiting the correlated nature of the variables within a data set.⁵³ Linear combinations of the correlated variables are derived such that the majority of the variance of the original data can be described by a few orthogonal components.⁴² PLS also relies on linear combinations of the original variables but differs from PCA in that it employs a least-squares regression analysis step to relate the extracted components to a response, in this case activation energies. Furthermore, in PLS the linear combinations of variables are chosen to maximize the correlation between the extracted components and the response.⁴³ The new components (called latent variables) capture the essential information of the original variables, but this approach can be applied without the typical drawbacks associated with using correlated descriptors in multiple linear regression.⁴⁴ MLR is the classical linear regression analysis based on using the original **X**-matrix and **Y**-vector to build an equation model which satisfies the least-squares principle.⁵⁴

All the generated statistical models used 15 geometric descriptors obtained from the QM/MM-optimized FAAH–oleamide complexes. These descriptors consisted of the key distances that change significantly during the course of the reaction. All descriptors were mean centered and scaled to unit variance because the numerical values of the descriptors vary significantly. This gives each variable the same opportunity of influencing the PCA/PLS model. The goodness of the correlation both in the PLS and MLR models can be reported using the coefficient of determination (R^2 , where values close to 1 indicate a good model) and the standard error. The greater the number of variables compared to observations in a model, the more likely it is that a strong chance correlation may be found and that the model is overfitted. To determine the significance of the models

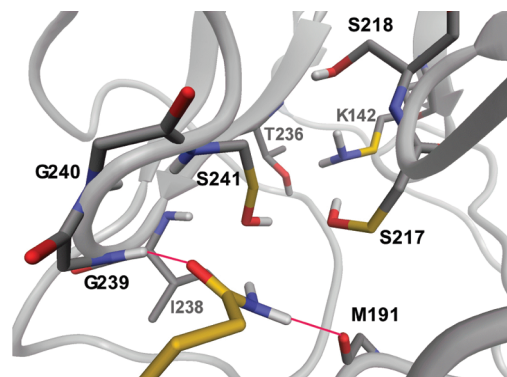


Figure 1. Representative QM/MM-optimized structure of the FAAH–oleamide complex (namely, snapshot 6 taken from the QM/MM MD simulation). Carbon atoms included in the quantum mechanical region are shown in yellow, and all the other carbons are shown in gray. Hydrogen bonds between OA and FAAH are also indicated.

generated, we applied the leave-one-out (LOO) cross-validation procedure,⁵⁵ where one observation is left out of the model-building process and its activation barrier is subsequently predicted by the reduced observation model. This is repeated until every observation has been left out and predicted at least once. A cross-validated regression coefficient, Q^2 , can then be calculated and compared to R^2 . The closer the R^2 value to the Q^2 value, the greater the confidence in the model one can have. A PCA/PLS component was considered significant if the Q^2 value was larger than 0.

PCA and PLS studies were performed with the SIMCA-P+ software.⁵⁶ MLR calculations were performed with Microsoft Excel97, employing the built-in statistical functions and automated macro procedures to determine the empirical value for regression coefficients and statistical parameters.

Results and Discussion

General Description of the Reaction. Multiple structures of the FAAH–OA enzyme–substrate complex were obtained from a QM/MM MD simulation (Supporting Information) based on the crystallographic coordinates of rat FAAH.⁵⁷ The complexes extracted from the MD trajectory showed an arrangement of the reactants similar to that previously published by our group⁵⁸ and other groups^{35,59} (Figure 1). The long lipophilic chain of the OA substrate was accommodated in the acyl chain binding site, while its polar head pointed toward the cytosolic access channel of the enzyme. The amide group shows polar interactions within the FAAH active site: while the carbonyl oxygen was anchored to the oxyanion hole, forming a hydrogen bond with Gly239, the oleamide NH group formed an additional hydrogen bond with the backbone CO group of Met191. With few exceptions (vide infra), the members of the catalytic triad were found to be well oriented to initiate the catalytic process, with the bridging Ser217 accepting a hydrogen bond from Ser241 and donating another one to Lys142. Lys142 formed two further hydrogen bonds, one with Ser218 and the other with Thr236, suggesting that, in agreement with mutagenesis,⁶⁰ Ser218 is important for catalysis.

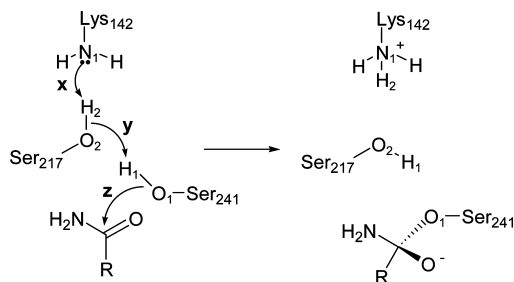


Figure 2. First step of oleamide hydrolysis catalyzed by FAAH: activation of Ser241 (X, Y) and formation of the tetrahedral adduct (Z). Labels are consistent with the definition of the reaction coordinates.

Starting from each structure of the enzyme–substrate complex, the first step of the acylation reaction was simulated by applying an adiabatic mapping strategy based on a hybrid PM3-CHARMM22 potential (Supporting Information). This approach has been used successfully in previous studies of the FAAH-catalyzed reaction involving amide³⁴ and ester⁵⁸ substrates and a carbamate inhibitor.⁶¹

Mutagenesis and kinetic experiments⁶² in conjunction with simulations^{15,34,35,58} suggested that the probable mechanism of the first step of the acylation reaction involves the

activation of Ser241 through the abstraction of a proton by the neutral Lys142, via the bridging residue Ser217 (X, Y), and the nucleophilic addition of Ser241 to the carbonyl carbon of oleamide (Z), generating the tetrahedral complex (Figure 2).³⁴

PESs were calculated by restraining two reaction coordinates, named R_x and R_y , which account for all the key chemical processes. The first coordinate $R_x = d[\text{O}_2, \text{H}_1] - d[\text{O}_1, \text{H}_1] - d[\text{O}_1, \text{C}]$ includes the proton abstraction from Ser241 performed by Ser217 (Y) and the nucleophilic attack (Z) led by Ser241 on the carbonyl carbon C. The second reaction coordinate $R_y = d[\text{O}_2, \text{H}_2] - [N_1, \text{H}_2]$ describes the proton transfer between Ser217 and the neutral Lys142 (X). A detailed description of the calculations is reported in the Supporting Information.

The mechanism is the same for all the FAAH–OA snapshots. The shape of the PESs and the relative positions of the stationary points are not affected by the geometry of the starting points. As examples, four PM3-CHARMM22 PESs are shown in Figure 3.

The first step of the acylation reaction can be described as follows: neutral Lys142 initiates reaction by accepting a proton from Ser217, leading to the formation of **2**. In turn,

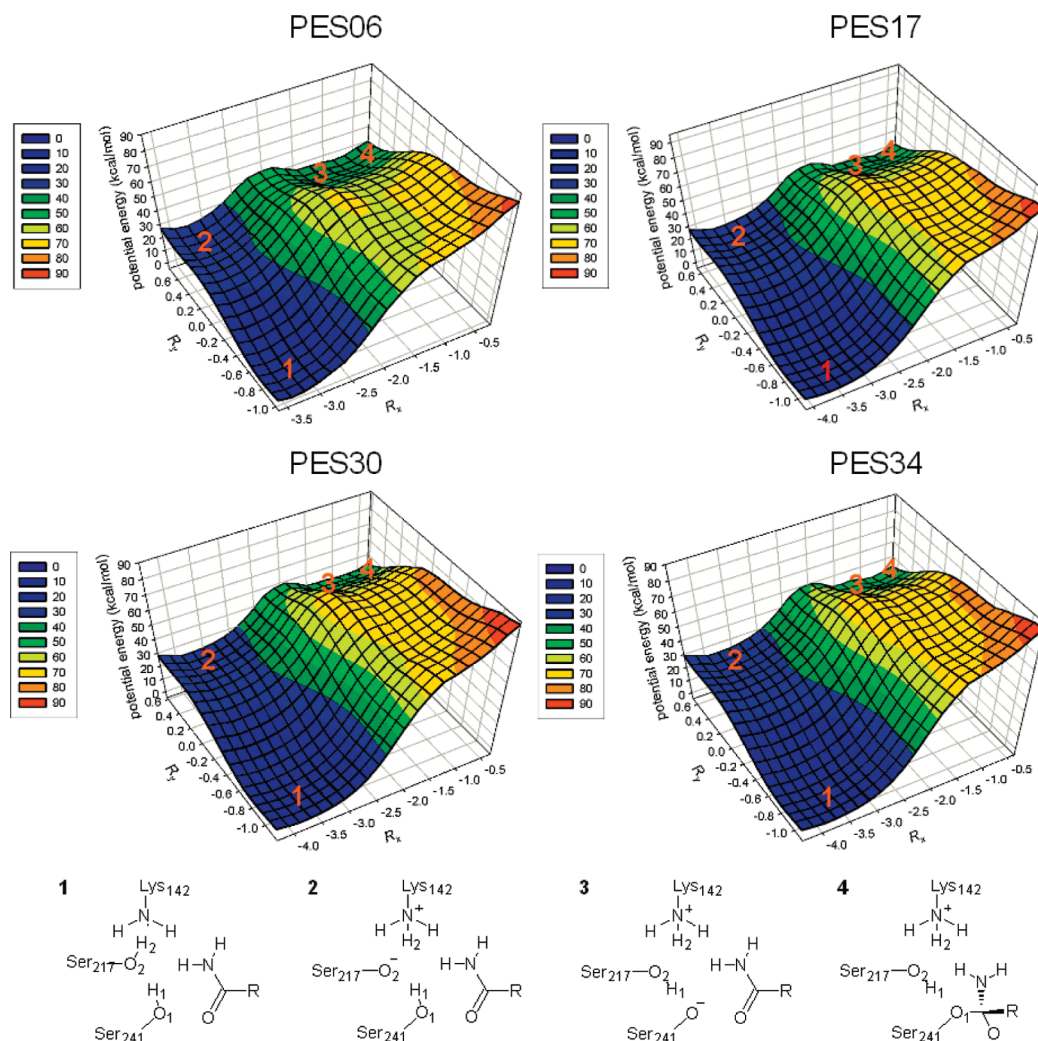


Figure 3. Representative PESs from four distinct snapshots (namely, 6, 17, 30, and 34) taken from QM/MM MD simulation. 1–4 are the minimum structures identified along the reaction pathway.

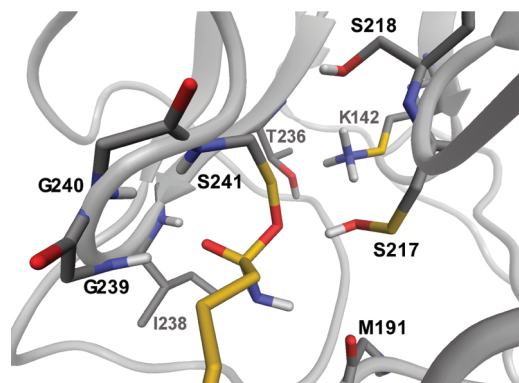


Figure 4. Representative TI of the FAAH–oleamide system (generated from snapshot 6). Bonds involving carbon atoms treated by quantum mechanics are shown in yellow, and all the other carbon atoms are shown in gray.

Ser217 deprotonates Ser241 to form the nucleophilic Ser241 anion (**3**). This short-lived intermediate **3** prompts the subsequent nucleophilic attack, leading to the tetrahedral intermediate (TI) **4**. The PESs indicate that the highest barrier is for deprotonation of Ser241. The absence of a significant (between 0.5 and 1 kcal mol⁻¹) barrier between **3** and **4** indicates that TI formation and Ser241 deprotonation are effectively concerted.

Table 1. Energy Barriers, Absolute Energies, and Geometrical Descriptors Taken from FAAH–Oleamide Complexes Employed in the Multivariate Analysis

snapshot	E_{att}^a	energy ^a	I238 ^b	G239 ^b	G240 ^b	S241 ^b	M191 ^b	S218 ^b	T236 ^b	CO ^b	O1C ^b	N1O1 ^b	H1O1 ^b	H2O2 ^b	H1O2 ^b	H2N1 ^b	N1O2O1 ^c
1	45.97	-1068.67	3.158	1.913	3.429	2.819	1.983	2.173	2.001	1.228	3.171	4.504	0.950	0.951	2.198	2.050	98.605
2	53.66	-1440.92	2.047	1.811	3.326	2.549	1.966	2.244	2.088	1.234	2.877	3.566	0.953	0.951	3.920	2.539	57.316
3	46.47	-1218.24	2.592	1.815	3.280	2.589	1.932	2.196	2.015	1.230	2.432	4.707	0.943	0.952	1.987	1.968	109.553
4	45.43	-1253.98	2.793	1.831	3.310	2.714	2.041	2.178	1.987	1.230	2.461	4.778	0.943	0.951	1.988	1.958	111.174
5	47.06	-1231.37	3.034	1.868	3.322	2.702	2.012	2.219	2.019	1.228	2.977	4.504	0.946	0.951	2.113	2.006	99.056
6	43.08	-1282.47	2.808	1.823	3.453	2.819	1.959	2.209	1.996	1.228	2.554	4.591	0.944	0.951	2.017	1.994	104.389
7	45.17	-1119.30	2.905	1.849	3.254	2.659	1.983	2.166	1.982	1.229	2.908	4.664	0.944	0.950	2.053	2.098	104.952
8	49.61	-1198.66	3.015	1.847	3.270	2.701	1.983	2.163	2.017	1.229	3.111	4.589	0.946	0.949	2.116	2.025	103.042
9	49.51	-1135.02	3.049	1.854	3.071	2.690	1.945	2.176	2.042	1.229	2.946	4.619	0.947	0.950	2.124	2.038	103.297
10	50.37	-1250.90	3.026	1.868	3.269	2.674	2.064	2.137	2.022	1.228	2.956	4.687	0.946	0.951	2.103	2.215	102.695
11	46.48	-1301.06	3.178	1.928	3.433	2.730	1.991	2.160	2.042	1.227	3.058	4.487	0.949	0.951	2.151	2.090	96.908
12	48.68	-1266.21	3.198	1.927	3.310	2.722	1.998	2.148	2.028	1.227	3.158	4.558	0.949	0.952	2.147	2.125	98.346
13	44.35	-1160.14	2.159	1.787	3.449	2.677	1.916	2.120	2.001	1.232	2.430	4.748	0.948	0.951	2.100	2.103	107.314
14	48.28	-1277.85	3.029	1.920	3.323	2.730	1.952	2.124	2.010	1.227	2.952	4.608	0.949	0.952	2.158	2.192	98.033
15	44.35	-1141.59	2.469	1.805	3.363	2.719	1.923	2.110	1.994	1.230	2.443	4.805	0.943	0.951	1.982	2.011	111.551
16	44.55	-1254.56	2.614	1.823	3.318	2.664	1.968	2.117	2.008	1.230	2.449	4.904	0.944	0.951	2.007	2.016	113.395
17	46.06	-1243.52	3.010	1.850	3.284	2.721	2.033	2.123	1.975	1.228	2.958	4.514	0.946	0.951	2.108	2.072	98.241
18	48.28	-1115.84	3.043	1.873	3.354	2.854	2.148	2.109	1.972	1.229	3.037	4.664	0.949	0.951	2.191	2.273	98.862
19	46.83	-1093.53	2.926	1.850	3.279	2.678	2.001	2.152	2.022	1.229	2.969	4.371	0.946	0.951	2.112	2.074	94.078
20	44.03	-1263.99	2.542	1.805	3.234	2.606	1.995	2.166	2.044	1.231	2.423	4.709	0.942	0.951	1.981	1.977	108.280
21	48.27	-1102.41	2.958	1.852	3.280	2.733	2.099	2.176	2.071	1.229	2.987	4.582	0.947	0.950	2.118	2.029	100.490
22	47.34	-1059.90	3.069	1.870	3.280	2.774	2.108	2.219	2.018	1.229	3.021	4.422	0.947	0.951	2.114	2.015	96.461
23	49.23	-1066.58	3.214	1.911	3.325	2.809	1.899	2.131	2.013	1.228	3.078	4.797	0.947	0.951	2.630	2.271	98.566
24	48.71	-1109.99	3.150	1.900	3.392	2.816	2.108	2.115	2.000	1.227	3.121	4.587	0.948	0.951	2.188	2.104	98.657
25	52.28	-1180.44	3.202	1.911	3.224	2.821	2.074	2.188	2.048	1.229	3.351	4.382	0.948	0.952	2.173	2.151	92.389
26	50.76	-1083.32	3.120	1.889	3.301	2.784	2.060	2.147	2.011	1.228	3.184	4.477	0.947	0.952	2.131	2.134	95.874
27	46.41	-1275.03	2.712	1.819	3.239	2.648	1.978	2.132	2.025	1.230	2.669	4.732	0.945	0.951	2.035	2.070	106.764
28	49.95	-1228.80	3.061	1.873	3.302	2.732	2.025	2.154	2.029	1.229	3.046	4.363	0.947	0.951	2.074	2.092	94.544
29	50.05	-1078.36	3.157	1.894	3.309	2.753	2.030	2.185	2.033	1.228	3.120	4.355	0.948	0.951	2.098	2.068	94.382
30	48.69	-1059.09	3.099	1.892	3.371	2.781	1.979	2.166	2.005	1.228	3.097	4.411	0.948	0.952	2.148	2.126	93.797
31	48.25	-1053.67	3.177	1.925	3.416	2.801	1.937	2.153	2.012	1.228	3.165	4.420	0.948	0.952	2.163	2.112	93.955
32	50.10	-1006.53	3.247	1.941	3.420	2.839	1.960	2.188	2.060	1.227	3.201	4.351	0.949	0.952	2.179	2.084	92.384
33	49.36	-1118.60	3.133	1.898	3.329	2.750	2.038	2.179	2.043	1.228	3.138	4.313	0.948	0.951	2.148	2.053	92.505
34	49.70	-1117.11	3.135	1.885	3.292	2.747	2.047	2.194	2.070	1.228	3.136	4.327	0.949	0.951	2.162	2.070	92.449
35	49.22	-1148.48	3.041	1.868	3.267	2.713	2.088	2.135	2.024	1.229	3.059	4.499	0.948	0.952	2.117	2.108	96.923
36	52.19	-1260.25	3.124	1.889	3.217	2.714	2.417	2.181	2.039	1.231	3.144	4.189	0.960	0.972	1.785	1.776	100.373
av	48.02	-1174.07	2.950	1.868	3.321	2.729	2.018	2.162	2.021	1.229	2.924	4.522	0.947	0.952	2.162	2.086	98.878
SD	2.50	95.21	0.289	0.041	0.060	0.071	0.091	0.033	0.027	0.001	0.268	0.233	0.003	0.004	0.325	0.119	9.276

^a Energies in kilocalories per mole. ^b Distances in angstroms. ^c Angles in degrees.

Visual inspection of the structures along points **3** and **4** showed that, as the nucleophilic attack proceeds, the hybridization of the reacting carbonyl carbon changes from sp² to sp³ and the distances between the carbonyl oxygen and the NH groups of the residues forming the oxyanion hole become smaller, improving the stabilization of the TI. The resulting protonated Lys142 is caged in a tight network of hydrogen bonds involving the proton shuttle Ser217 and two peripheral residues of the active site, namely, Thr236 and Ser218 (Figure 4).

PCA and PLS Analysis. The activation barriers, extracted from the PESs, are given in Table 1 for the 36 snapshots. In the same table, geometric parameters (15 variables total) for the FAAH–OA structures employed for the statistical analysis are also reported. These descriptors were selected on the basis of interactions observed in the enzyme–substrate complex and in the TI (Figures 1 and 4). They include distances between the carbonyl oxygen of OA and the N–H backbone groups belonging to the oxyanion hole residues (I238, G239, G240, S241), the N–H of OA and the amide binding site (M191), the Lys142 polar hydrogens and their environment (S218, T236), and the reactant atoms such as the carbonyl double bond distance CO, the nucleophile attacking distance O1C, and others (N1O1, H1O1, H2O2,

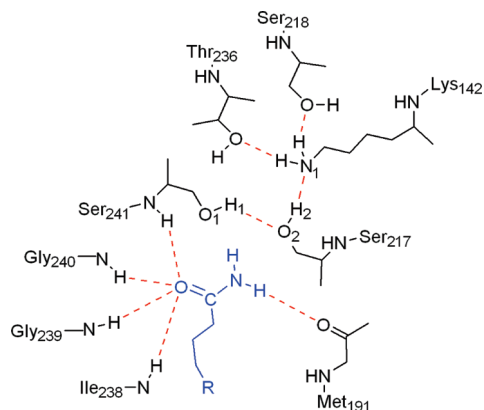


Figure 5. Two-dimensional representation of oleamide (in blue) within the FAAH active site (black line). Polar interactions (hydrogen bonds) are highlighted with red dotted lines. Atom labels are consistent with the reaction coordinate as well as with variable names included in the statistical analysis.

H1O2, H2N1; see Figure 5). The angle formed by Lys142 nitrogen, Ser217 oxygen, and Ser241 oxygen (N1O2O1) was also considered. In the same table the mean values of the collected variables are reported, together with their standard deviations (SDs).

The same geometrical descriptors were also collected for the TS structures identified on the PESs (Table S1, Supporting Information). Comparison of the standard deviations indicates that the differences between the enzyme–substrate complex geometries are in general larger than those between the corresponding transition-state structures, with a few exceptions (H1O1 and H2O2 descriptors). This also suggests that the accessible conformational space of the FAAH–OA system is wider at the enzyme–substrate complex than at the TS, at least along the specified variables. For these reasons, multivariate analyses were limited to the first family of structures, i.e., the enzyme–substrate complexes.

PCA was applied to the geometrical descriptors reported in Table 1. PCA produces a summary of the data structure showing how the observations (here the snapshots) are related to each other and highlights any deviating (outlier) observations. The advantage of PCA is that the information contained in the original \mathbf{X} -matrix (here constituting 15 descriptors) can be concentrated in 2–3 principal components, representing truly independent effects.⁴² Here, the first two principal components (called $t[1]$ and $t[2]$) describe approximately 58% of the total variance in the 15 descriptors. The model has a negative cross-validated Q^2 , indicating that some inconsistency may be present in the data set.⁴⁶ Indeed, the score plot of the first two components reported in Figure 6 indicates the presence of one outlier, namely, snapshot 2, which lies outside Hotelling's T^2 confidence ellipse.⁶³

The analysis of the raw data suggests that snapshot 2 differs from the other snapshots in that Ser241 (H1) does not form a hydrogen bond with Ser217 (O2) (H1O2 = 3.920 Å), an interaction which is crucial for proton transfer between them. Visual inspection of the FAAH–OA complex showed that, in snapshot 2, Ser217 adopts a different conformation at the binding site, with χ_1 in an *anti* conformation rather than in the *gauche* (+) conformation as observed for all other

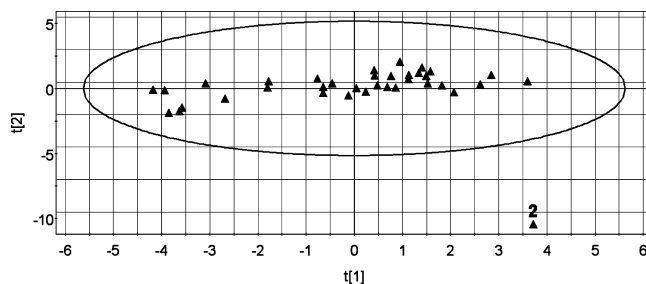


Figure 6. PCA score plot of component 1, $t[1]$, against component 2, $t[2]$. The curve displayed is Hotelling's ellipsoid, describing the area in which an observation can be expected to fall with a probability of 95%.

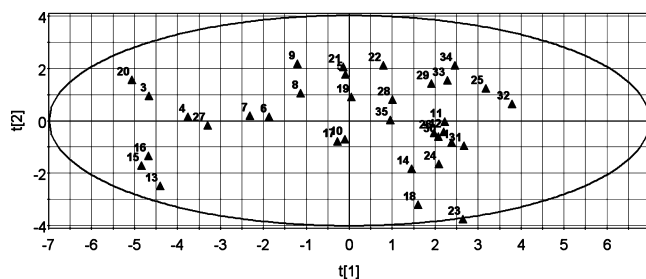


Figure 7. PCA score plot of component 1, $t[1]$, against component 2, $t[2]$, for the second data set (snapshots 2 and 36 excluded).

starting geometries. This difference is also mirrored by a N1O2O1 value of 57.31°, far from the mean value of the series for this angle (mean \pm SD = 98.88 \pm 9.28°). Snapshot 2 was thus excluded from the data set as further multivariate analysis would essentially only explain the differences between this structure and all the others. To detect more moderate outliers, which are not powerful enough to shift the model plane and therefore to show up as outliers in score plots, we looked at the observation distance to the generated model in \mathbf{X} -space, calculating DmodX, also known as the residual error.⁴⁶ Snapshot 36 can be identified as a moderate outlier and was also excluded from further analysis. In this case (snapshot 36), analysis of the raw data suggests a different accommodation for the $-\text{NH}_2$ group of OA, as the interaction with the carbonyl oxygen of Met191 (M191 = 2.417 Å) is not optimal compared to that observed in all other snapshots (mean value \pm SD = 2.018 \pm 0.061 Å). A final PCA was performed with a data set composed of the remaining 34 observations and 15 variables ($N = 34$; $X = 15$). The observations, reported in a new score plot (Figure 7), are more evenly distributed when compared to the previous data set, indicating a more homogeneous \mathbf{X} -matrix.

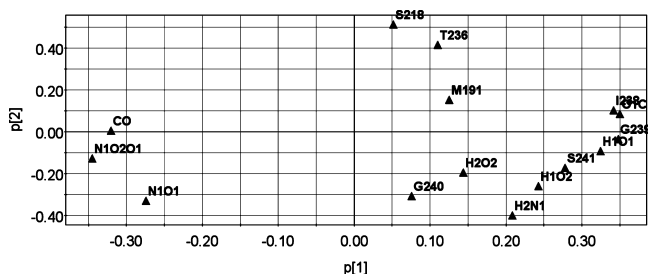
The performance of the PCA is reported in Table 2. The first component explains 48% of the total variance, while the second one extracts an additional 16% of information. The introduction of other components is not justified as both the third and fourth PCs have a negative Q^2 . Thus, the best model has a cumulative (cum) R^2 of 0.64 with a positive cross-validated Q^2 (cum) of 0.40, indicating that descriptive and predictive powers are satisfactory to perform further statistical analysis.

The loadings plot (Figure 8) allows interpreting the score plot further, as it displays the relationship among the

Table 2. Summary of the PCA Analysis^a

comp no.	$R^2(\mathbf{X})$	$R^2(\mathbf{X})(\text{cum})$	Q^2	$Q^2(\text{cum})$
1	0.480	0.480	0.387	0.387
2	0.158	0.638	0.025	0.402
3	0.093	0.730	-0.075	0.357
4	0.071	0.802	-0.140	0.292

^a $R^2(\mathbf{X})$ refers to the variance in the descriptor matrix described by the PCA models. Q^2 refers to the cross-validated $R^2(\mathbf{X})$ for the PCA model. The comp no. is the number of the respective principal component.

**Figure 8.** PCA loading plot of the first two PCs of the data set where snapshots 2 and 36 were excluded.

variables.⁴² The loadings are the weights of the original variables when calculating the principal component scores. Score and loading plots are complementary; thus, a pattern observed in the score plot can be interpreted by looking along that direction in the loading plot. On the loading plot, descriptors found close to each other in both dimensions are highly correlated (e.g., I238 and O1C). Those that are similar on just one component are still correlated, but to a lesser extent (e.g., S218 and G240), whereas those at opposite ends of the components are inversely correlated (O1C and CO). Descriptors found at the extreme ends of the x or y axes have the most significant impact on the component that defines that axis, whereas unimportant variables are around the origin of the plot. Here, many of the descriptors have large positive or negative loadings (~ 0.3) on the x axis, indicating that, as expected, the descriptors in the data set are highly correlated.

The first principal component encodes information mainly related to the accommodation of the amide group of OA into the FAAH active site. The most important variables are (i) with positive loadings, O1C (the nucleophile attacking distance) and I238 and G239 distances, describing two (of four) hydrogen bonds formed by the carbonyl oxygen with the oxyanion hole, and (ii) with negative loadings, the N1O2O1 angle and the CO distance, measuring the length of the carbonyl double bond of oleamide. These five variables were also highly correlated, with absolute values of their correlation coefficients often higher than 0.8 and never lower than 0.65 (Table S2, Supporting Information). Also S241, the third anchor point in the oxyanion hole, N1O1, the distance between the Lys142 nitrogen and the nucleophile oxygen, and H1O1 showed large loadings on the first component, but they showed slightly lower correlation coefficients with the variables discussed above and with each other (Table S2). Thus, these eight variables describe a concerted movement within the active site, resulting in deeper accommodation of OA into the oxyanion hole, approaching the nucleophile and polarization of the carbonyl bond. In

Table 3. Summary of the PLS Analysis^a

comp no.	$R^2(\mathbf{X})$	$R^2(\mathbf{X})(\text{cum})$	$R^2(\mathbf{Y})$	$R^2(\mathbf{Y})(\text{cum})$	Q^2	$Q^2(\text{cum})$
1	0.473	0.473	0.628	0.628	0.582	0.582
2	0.114	0.587	0.148	0.776	0.219	0.674
3	0.124	0.711	0.019	0.795	-0.128	0.641
4	0.061	0.772	0.014	0.809	-0.374	0.605

^a $R^2(\mathbf{X})$ refers to the variance in the descriptor matrix described by the PLS models. $R^2(\mathbf{Y})$ refers to the variance in the energy explained by the PLS model. Q^2 refers to the cross-validated $R^2(\mathbf{Y})$ for the PLS model. The comp no. refers to the component number.

fact, shorter hydrogen bonds in the oxyanion hole (I238, G239, S241) correspond to shorter O1C distances and to a longer and possibly more polarized carbonyl bond. As a consequence, the distance O1C can be considered representative of all these concerted movements.

Looking at the score plot, it is possible to note the presence of two observation groups with respect to the O1C descriptor: all the snapshots with high O1C distances (e.g., longer than 2.95 Å) have positive $t[1]$ and are located on the right-hand side of the plot, whereas those with smaller O1C values are on the left. The second component explains 16% of the total variation in the descriptors and is dominated by the positive effect of S218 and T236, accounting for the hydrogen bond network stabilizing the Lys142 side chain nitrogen. It is interesting to observe that these two variables present low intercorrelation ($r = 0.43$), and they are barely correlated with other variables having significant loadings on the second component, such as H2N1 and G240 (see Table S2, Supporting Information). These variables seem therefore to bring a common portion of information, but their variation is much less concerted than that described by the first component.

PLS is a regression technique for modeling the relationship between multiple X and Y variables, useful in the case of a complex \mathbf{X} -descriptor matrix, where traditional MLR cannot be applied due to the presence of highly correlated variables.⁴³ PLS tries to extract only statistically relevant information contained in the \mathbf{X} -matrix (in that way it is similar to PCA) and to use such information to build regression models.⁴³ This means that, in PLS, a small group of new independent variables (called “latent variables”) will be generated to explain the \mathbf{Y} -space. In the current investigation, the activation energy (E_{att}) was used as the response variable. It changes significantly among the 36 observations, with an absolute range of variation of ~ 11 kcal mol⁻¹, corresponding to 22% of the E_{att} mean value. An exploratory regression analysis reveals that E_{att} does not depend on the absolute energy of the system (data not shown). This indicates that the QM/MM protocol applied is robust with respect to small changes in the composition of the system.

Table 3 reports the PLS models for the data set consisting of 34 observations, 1 response variable, and 15 descriptors ($N = 34$; $Y = 1$; $X = 15$) where the overall R^2 and Q^2 statistics change as a function of the model complexity. The first latent variable describes 47% ($R^2(\mathbf{X})$) of the information contained in the \mathbf{X} -matrix and correlates with \mathbf{Y} , giving an $R^2(\mathbf{Y})$ of 0.63 and Q^2 of 0.58. The introduction of a second latent variable slightly improves the model, as both the

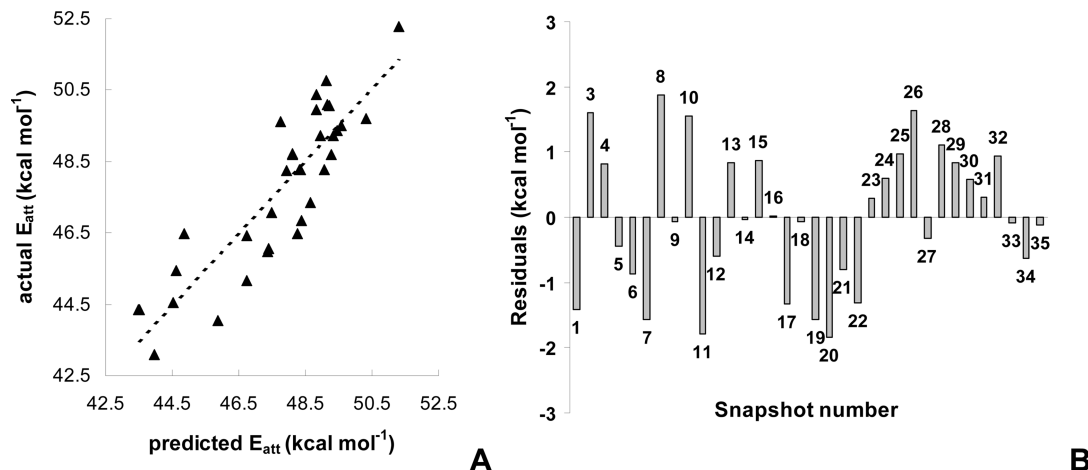


Figure 9. Observed E_{att} values vs those calculated by the PLS model (A, left) and residuals, calculated as the difference between the actual E_{att} and calculated E_{att} (B, right).

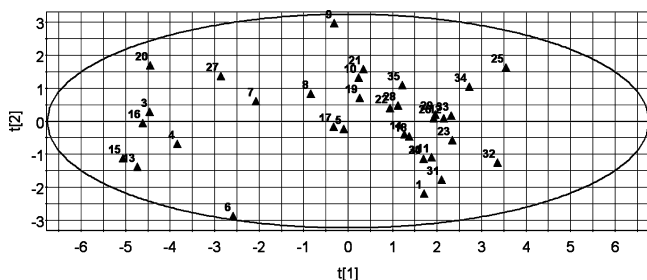


Figure 10. PLS score plot of the two first latent variables of the data set.

cumulative $R^2(Y)$ and Q^2 increase to values of 0.78 and 0.67, respectively. Here, two components appear appropriate as Q^2 starts to decrease when additional latent variables are introduced.

The performance of this two-variable model is fairly good, with 78% of the variation in E_{att} explained by the PLS equation, giving an acceptable root-mean-square error (RMSE) of 1.098 kcal mol⁻¹. Analysis of the residuals (the difference between the actual E_{att} and calculated E_{att} , Figure 9) indicates that some snapshots (e.g., 3, 8, 10, 11, 20, and 26) are poorly predicted. However, these structures are homogeneous observations in the PCA space (Figure 7); therefore, we chose to retain them in the present analysis.

The score plot derived from the 15-descriptor PLS models (Figure 10) does not differ significantly from that derived from the PCA model built using the same data set. The presence of two groups of observations can be detected in this case as also discussed for PCA, particularly along the first latent variable ($t[1]$). The similarity and differences between the PLS and PCA components can be understood by an analysis of the PLS weight plot (W^*C plot, Figure 11), corresponding to the loading plot in PCA. However, the weight plot in PLS differs somewhat from the loading plot in PCA in that the response variable (E_{att}) is also displayed. It is thus possible to determine the effect of a particular descriptor on E_{att} on the basis of their spatial relationship in the PLS weight plot.⁴³

The most important descriptors for the first latent variable, positively correlated with E_{att} , are the nucleophile attacking distance O1C and I238 distance, describing the polar

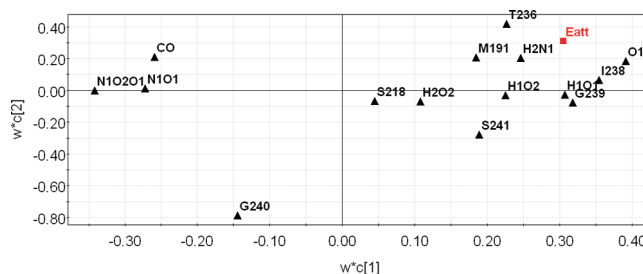


Figure 11. W^*C weight plot for latent variables 1 and 2. The Y variable (E_{att}) is marked in red. This plot shows both the X weights (W^*) and Y weights (C), and thus the correlation between the original X-matrix of geometrical descriptors and Y (E_{att}).

interaction between the Ile238 NH backbone and the carbonyl oxygen of OA. On the other hand, the N1O2O1 variable, describing the angle formed among the Lys142 basic nitrogen, Ser217 hydroxyl oxygen, and Ser241 hydroxyl oxygen is negatively correlated with the activation barrier. (e.g., the higher the angle value, the lower the barrier).

With respect to PCA, some differences also arise from the second latent variable, which is now dominated by G240, describing the polar interaction distance between the OA carbonyl oxygen and the NH backbone of the oxyanion hole residue Gly240, and to a lesser extent by T236, describing the hydrogen bond distance between the Thr236 hydroxyl hydrogen and Lys142 side chain nitrogen. Finally, compared to PCA, S218 loses its influence on the model, being located close to the origin of the W^*C plot.

A complementary way to recognize the impact of the original variables on the PLS model is to calculate the VIP (variable importance in the projection), summarizing the importance of the X variables in both the descriptor and response spaces. VIPs are squared functions of the PLS weights, W^* , taking into account the amount of explained Y variance in each dimension. Descriptors with a VIP larger than 1 (e.g., O1C, G240) are the most influential for modeling the observed variation in E_{att} (Figure 12), consistent with what was observed in the W^*C weight plot for latent variables 1 and 2 (Figure 11).

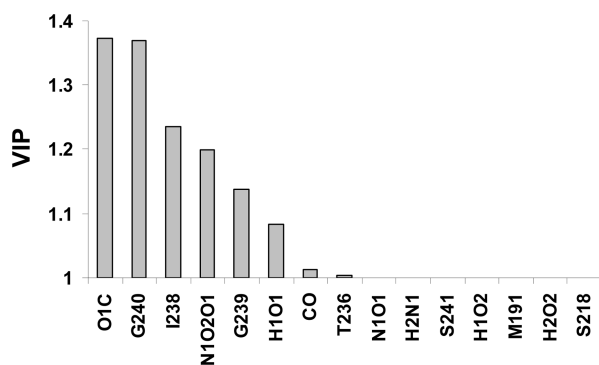


Figure 12. VIP values calculated for the descriptors employed in the PLS analysis. To highlight the most significant variables, the VIP scale starts from 1.

In our previous investigation on conformational fluctuations in FAAH catalysis, we identified N101 as the crucial determinant for reactivity.¹⁵ With the current data set, this descriptor plays a negligible role in affecting the barrier of the acylation (shown by its small VIP value (<1)). This apparent discrepancy appears to be due to the different method employed to generate the enzyme–substrate complexes. Indeed, inclusion of these structures (renamed snapshots 37–40) in the principal component analysis ($N = 38$; $X = 15$) shows that these enzyme–substrate complexes are outliers, placed at the border of Hotelling’s T^2 confidence ellipsoid (Figure S3, Supporting Information). PCA reveals that the main sources of variation between snapshots 37–40 and all the others (Figure S4, Supporting Information) reside both in accommodation of the amide group of OA in the FAAH active site (as revealed by the loading values of G239 and O1C along the first component) and in bond distances involving hydrogen atoms crucial for the reaction (e.g., H102, H2O2, H2N1 for the first component and H101 for the second one).

MLR Analysis and Chemical Interpretation of the Models. The interpretation of the latent variables generated in a PLS regression may be difficult, due to the necessity to combine the original descriptors to form derived variables.⁶⁴

An MLR analysis was therefore also performed, employing the data set composed of the 15 original descriptors and 34 homogeneous observations ($N = 34$; $Y = 1$; $X = 15$), exploring linear relationships between the response data and smaller subsets of descriptors. However, it should be considered that MLR presents a higher risk of oversimplification, giving emphasis on a specific variable, or overfitting, by inclusion of noisy variables.

Among the 15 possible univariate models, O1C gives the best correlation with E_{att} :

$$E_{\text{att}} = (28.369 \pm 2.550) + (6.601 \pm 0.865)\text{O1C} \quad (1)$$

$$n = 34 \quad R^2 = 0.646 \quad s = 1.354$$

$$F = 58.1 \quad (p = 1.1 \times 10^{-8}) \quad Q^2 = 0.602$$

The positive relationship between E_{att} and O1C indicates that the activation barrier is lowered if the nucleophilic oxygen of Ser241 and the carbonyl carbon of OA are in close proximity. However, current and previous³⁴ calculations

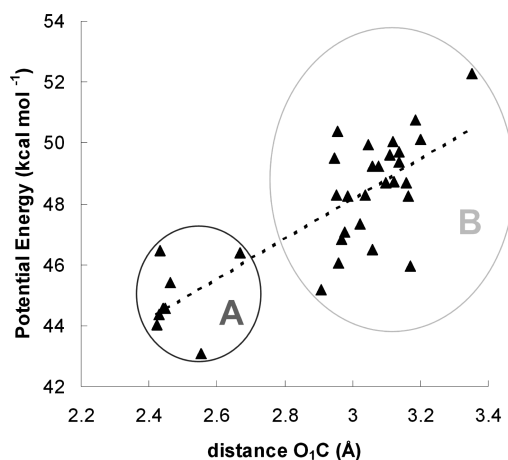


Figure 13. Calculated activation energy E_{att} (kcal mol⁻¹) vs the nucleophile attack distance O1C (Å).

indicated that the TS for the first step of the acylation corresponded to the deprotonation of Ser241 rather than to the nucleophilic attack. It is likely that the closeness of O₁ and C, induced by a coming together of the substrate with the S241 nucleophile oxygen, facilitates the key proton transfer of the reaction. It is possible that the more the substrate is pushed toward S241, the more it resembles the TS structure, leading to a lower activation barrier (vide infra).

The scatter plot of E_{att} vs O1C shows the presence of two different clusters in the data set (Figure 13). All the starting structures where the O1C distance is shorter than 2.8 Å belong to cluster A, as is the case for snapshots 3, 4, 6, 13, 15, 16, 20, and 27. These generally have a lower energy barrier than members of cluster B (where the O1C distance exceeds 2.8 Å), which contains the remaining 26 observations.

Although this simple linear regression mainly explains the difference in reactivity between these two clusters rather than highlighting differences among structures within any single group, it gives important insights for the reaction. Visual inspection of the 34 TS structures for all the calculated PESs (with a proton being transferred from Ser241 to Ser217) shows that, on average, the O1C distance measures 2.035 ± 0.025 Å (mean \pm SD). Conformational fluctuations of the OA carbonyl in the FAAH active site influence the catalytic process: in the substrate complex, a O1C distance close to that in the TS structure is associated with a lower activation barrier.

As already discussed, O1C is significantly correlated with other geometrical descriptors describing a concerted process. In fact, other related variables gave significant linear correlations with E_{att} , even if with lower descriptive and predictive power.

PLS analysis indicated the presence of another effect accounting for the differences in the calculated barriers. Among all the possible 105 bivariate linear regression equations, only two models have satisfactory descriptive and predictive power, bringing independent information (determinant of the correlation matrix >0.9), namely, eqs 2 and 3. In both cases, the nucleophilic attack distance O1C remains an important variable for modeling the variance of the activation barrier: its regression coefficient does not change significantly from one equation to the others.

$$E_{\text{att}} = (55.838 \pm 9.379) + (6.564 \pm 0.773)\text{O1C} - (8.250 \pm 2.733)\text{G240} \quad (2)$$

$$n = 34 \quad R^2 = 0.726 \quad s = 1.209 \\ F = 41 \quad (p = 2.0 \times 10^{-9}) \quad Q^2 = 0.674$$

$$E_{\text{att}} = (-13.571 \pm 18.538) + (5.99 \pm 0.857)\text{O1C} + (21.663 \pm 9.495)\text{T236} \quad (3)$$

$$n = 34 \quad R^2 = 0.696 \quad s = 1.274 \\ F = 35 \quad (p = 1.1 \times 10^{-8}) \quad Q^2 = 0.637$$

Equation 2 shows a negative relationship between E_{att} and G240. The Gly240 NH group is one of the anchor points (in the oxyanion hole) for the carbonyl oxygen of OA within the FAAH active site; thus, the G240 descriptor can be considered as an indirect measure of the tightness of binding to the substrate. The analysis indicates that the shortening of this distance in the substrate complex significantly increases the barrier heights for the reaction.

In eq 3, the positive correlation between E_{att} and T236 indicates that shortening of the hydrogen bond distance between Lys142 (acting as the hydrogen bond donor) and the Thr236 side chain oxygen (acting as the hydrogen bond acceptor) in the substrate (Michaelis) complex lowers the barrier height for the reaction. It could be speculated that a tighter interaction between these two residues would result in an indirect increase of the nitrogen basicity of Lys142. Such a change would accelerate the reaction, consistent with the role of Lys142, acting as a general base in the early phase of the catalysis.^{34–36}

As expected, the two variables included with O1C in eqs 2 and 3 are those having the largest absolute values in the second latent variable of the PLS model, with consistent signs. On the other hand, the low correlation between the variables G240 and T236 ($r = -0.27$) indicates that they contain additional effects that may introduce noise in the MLR models, as well as the common information leading to their projection into the latent variable. In fact, the coefficient/standard error ratio for T236 in eq 2 is only barely significant in a t test ($p = 0.030$), while the coefficient of G240 in eq 2 is more significant ($p = 0.005$).

It is important to recognize that structural fluctuations associated with reaction could take place in any environment and so are not necessarily of any catalytic benefit. Conformational effects of the type identified here (e.g., fluctuations of the O1C distance) could be observed for a comparable (“reference”) reaction in solution. In such a case, the fluctuation itself would not be necessarily a source of catalysis,⁶ but part of the “complete” reaction coordinate, able to better describe the whole catalytic process. In enzymes where reaction may take place via distinct, high-energy conformations, definition of the entire reaction coordinate may be important: it is possible that the ability to efficiently undergo fluctuations leading to reactive conformations may be associated with substrate specificity. In

any case, a full description of a reaction within an enzyme requires characterization of motions associated with the reaction.

Conclusions

We report here the first example of multivariate analysis of enzyme reactivity, applied to an extended data set of FAAH–oleamide complexes extracted from a QM/MM MD simulation. A total of 36 independent PESs for the first step of the acylation reaction were calculated and analyzed to investigate the relationship between the geometry of the starting structure and the barrier height of the reaction. A total of 15 geometrical descriptors were collected from the active site of the 36 starting structures and employed for multivariate PCA and PLS and MLR analysis to assess the impact of conformational fluctuations of the active site on the calculated activation barrier. The variable selection strategy was to some extent based on our knowledge of the FAAH–substrate system and its catalytic mechanism. However, some general insights resulting from the present analysis may be transferable to other catalytic systems, e.g., the importance of considering distances describing polar interactions between the enzyme and the substrate or between members of the catalytic residues.

Results from PCA indicated that different “families” of enzyme–substrate conformations arise from QM/MM MD simulation and that a rarely sampled conformational state (e.g., snapshot 2) can be identified along a trajectory. Sampling is a crucial point of enzyme reaction simulation, as an inappropriate choice of starting structure could lead to an incorrect mechanism being modeled. A proper equilibration of the starting structure is therefore required, and multivariate PCA could help in selecting starting points which are different from each other in the space of catalytically relevant geometrical parameters.

Once outliers are identified and excluded from the data set, the relationship between the activation barrier and geometrical parameters can be investigated. PLS analysis revealed the presence of two distinct and fully independent geometrical effects, explaining ~78% of the response variable (i.e., the barrier). MLR confirms the results obtained with PLS, indicating that conformational fluctuations associated with the nucleophile attacking distance (O1C), substrate binding (G240), or stabilization of the general base Lys142 (T236) play a significant role in determining the activation barrier. Conformational fluctuations of the active site affecting these crucial distances, also of a few fractions of an angstrom, may considerably influence the calculated reaction barrier, contributing up to 20% of the variation from the mean value.

From a technical point of view, our findings highlight the importance of using a large number of starting structures in calculations of potential energy surfaces for enzyme reactions and confirm that structural fluctuation is an essential part of the process of the reaction within the enzyme. Furthermore, the use of multiple starting structures can capture protein motions (usually neglected in energy-minimization studies of enzymatic processes), along the specified reaction coordinate, identifying catalytically relevant enzyme–substrate

complexes and giving a rough estimate of the energy required to preorganize the active site for catalysis (for instance, looking at the standard deviation of the calculated barriers).

The results warn against simplistic analyses of structural effects, because multiple subtle inter-related structural factors affect the barrier height and shape for reactions in enzymes. The use of multivariate statistical approaches is crucial for explaining quantitatively the differences in the calculated barrier for different starting structures. Detailed statistical analyses of the type applied here should be useful in analyses of other enzymes to mine potentially hidden conformational effects. This should be considered when the effects of protein dynamics on enzyme reactions are analyzed.

Acknowledgment. A.J.M. thanks the Engineering and Physical Sciences Research Council (EPSRC) for support: A.J.M. is an EPSRC Leadership Fellow. The Thai Government is also acknowledged for funding (J.S.).

Supporting Information Available: Computational details for QM/MM simulations, correlation matrix, and supplementary figures for PCA. This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- (1) Karplus, M.; Gao, Y. Q.; Ma, J.; van der Vaart, A.; Yang, W. Protein Structural Transitions and Their Functional Role. *Philos. Trans. R. Soc. London, A* **2005**, *363*, 331–355.
- (2) Min, W.; English, B. P.; Luo, G.; Cherayil, B. J.; Kou, S. C.; Xie, X. S. Fluctuating Enzymes: Lessons from Single-Molecule Studies. *Acc. Chem. Res.* **2005**, *38*, 923–931.
- (3) Qian, H.; Shi, P. Z. When Does the Michaelis-Menten Equation Hold for Fluctuating Enzymes? *J. Phys. Chem. B* **2009**, *113*, 2225–2230.
- (4) Gorfe, A. A.; Lu, B.; Yu, Z.; McCammon, J. A. Enzymatic Activity versus Structural Dynamics: The Case of Acetylcholinesterase Tetramer. *Biophys. J.* **2009**, *97*, 897–905.
- (5) Liu, Y. H.; Konermann, L. Conformational Dynamics of Free and Catalytically Active Thermolysin Are Indistinguishable by Hydrogen/Deuterium Exchange Mass Spectrometry. *Biochemistry* **2008**, *47*, 6342–6351.
- (6) Olsson, M. H.; Parson, W. W.; Warshel, A. Dynamical Contributions to Enzyme Catalysis: Critical Tests of a Popular Hypothesis. *Chem. Rev.* **2006**, *106*, 1737–1756.
- (7) Garcia-Viloca, M.; Gao, J.; Karplus, M.; Truhlar, D. G. How Enzymes Work: Analysis by Modern Rate Theory and Computer Simulations. *Science* **2004**, *303*, 186–195.
- (8) Olsson, M. H.; Mavri, J.; Warshel, A. Transition State Theory Can Be Used in Studies of Enzyme Catalysis: Lessons from Simulations of Tunnelling and Dynamical Effects in Lipoygenase and Other Systems. *Philos. Trans. R. Soc. London, B* **2006**, *361*, 1417–1432.
- (9) Claeysens, F.; Harvey, J. N.; Manby, F. R.; Mata, R. A.; Mulholland, A. J.; Ranaghan, K. E.; Schütz, M.; Thiel, S.; Thiel, W.; Werner, H. J. High-Accuracy Computation of Reaction Barriers in Enzymes. *Angew. Chem., Int. Ed.* **2006**, *45*, 6856–6859.
- (10) Honkala, K.; Hellman, A.; Remediakis, I. N.; Logadottir, A.; Carlsson, A.; Dahl, S.; Christensen, C. H.; Nørskov, J. K. Ammonia Synthesis from First-Principles Calculations. *Science* **2005**, *307*, 555–558.
- (11) Pentikäinen, U.; Pentikäinen, O. T.; Mulholland, A. J. Cooperative Symmetric to Asymmetric Conformational Transition of the Apo-Form of Scavenger Decapping Enzyme Revealed by Simulations. *Proteins* **2008**, *70*, 498–508.
- (12) Thorpe, I. F.; Brooks, C. L. Conformational Substates Modulate Hydride Transfer in Dihydrofolate Reductase. *J. Am. Chem. Soc.* **2005**, *127*, 12997–13006.
- (13) Karplus, M.; McCammon, J. A. Dynamics of Proteins: Elements and Function. *Annu. Rev. Biochem.* **1983**, *53*, 263–300.
- (14) Villà, J.; Warshel, A. Energetics and Dynamics of Enzymatic Reactions. *J. Phys. Chem. B* **2001**, *105*, 7887–7907.
- (15) Lodola, A.; Mor, M.; Zurek, J.; Tarzia, G.; Piomelli, D.; Harvey, J. N.; Mulholland, A. J. Conformational Effects in Enzyme Catalysis: Reaction via a High Energy Conformation in Fatty Acid Amide Hydrolase. *Biophys. J.* **2007**, *92*, L20–L22.
- (16) van der Kamp, M. W.; Mulholland, A. J. Computational Enzymology: Insight into Biological Catalysts from Modeling. *Nat. Prod. Rep.* **2008**, *25*, 1001–1014.
- (17) Cavalli, A.; Carloni, P.; Recanatini, M. Target-Related Applications of First Principles Quantum Chemical Methods in Drug Design. *Chem. Rev.* **2006**, *106*, 3497–3519.
- (18) Karplus, M.; McCammon, J. A. Molecular Dynamics Simulations of Biomolecules. *Nat. Struct. Biol.* **2002**, *9*, 646–652.
- (19) Senn, H. D.; Thiel, W. QM/MM Studies of Enzymes. *Curr. Opin. Chem. Biol.* **2007**, *11*, 182–187.
- (20) Warshel, A. Computer Simulations of Enzyme Catalysis: Methods, Progress, and Insights. *Annu. Rev. Biophys. Biomol. Struct.* **2003**, *32*, 425–443.
- (21) Friesner, R. A.; Guallar, V. *Ab Initio* Quantum Chemical and Mixed Quantum Mechanics/Molecular Mechanics (QM/MM) Methods for Studying Enzymatic Catalysis. *Annu. Rev. Phys. Chem.* **2005**, *56*, 389–427.
- (22) Field, M. J.; Bash, P. A.; Karplus, M. A Combined Quantum Mechanical and Molecular Mechanical Potential for Molecular Dynamics Simulations. *J. Comput. Chem.* **1990**, *11*, 700–733.
- (23) Senn, H. M.; Thiel, W. QM/MM Methods for Biomolecular Systems. *Angew. Chem., Int. Ed.* **2009**, *48*, 1198–1229.
- (24) Mulholland, A. J. Computational Enzymology: Modelling the Mechanisms of Biological Catalysts. *Biochem. Soc. Trans.* **2008**, *36*, 22–26.
- (25) Hu, H.; Yang, W. Free Energies of Chemical Reactions in Solution and in Enzymes with *ab Initio* Quantum Mechanics/Molecular Mechanics Methods. *Annu. Rev. Phys. Chem.* **2008**, *59*, 573–601.
- (26) Mulholland, A. J. The QM/MM Approach to Enzymatic Reactions. In *Theoretical Biochemistry*; Eriksson, L. A., Ed.; Elsevier: Amsterdam, 2001; pp 597–653.
- (27) Cui, Q.; Karplus, M. Quantum Mechanical/Molecular Mechanical Studies of the Triosephosphate Isomerase-Catalyzed Reaction: Verification of Methodology and Analysis of Reaction Mechanisms. *J. Phys. Chem. B* **2002**, *106*, 1678–1698.
- (28) Lodola, A.; Woods, C. J.; Mulholland, A. J. Applications and Advances of QM/MM Methods in Computational Enzymology. In *Annual Reports in Computational Chemistry*; Wheeler, R. A., Spellmeyer, D. C., Eds.; Elsevier: Amsterdam, 2008; Vol. 4, Chapter 9, pp 155–169.

- (29) Kamerlin, S. C.; Haranczyk, M.; Warshel, A. Progress in *ab Initio* QM/MM Free-Energy Simulations of Electrostatic Energies in Proteins: Accelerated QM/MM Studies of pKa, Redox Reactions and Solvation Free Energies. *J. Phys. Chem. B* **2009**, *113*, 1253–272.
- (30) Bowman, A. L.; Ridder, L.; Rietjens, I. M.; Vervoort, J.; Mulholland, A. J. Molecular Determinants of Xenobiotic Metabolism: QM/MM Simulation of the Conversion of 1-Chloro-2,4-dinitrobenzene Catalyzed by M1-1 Glutathione S-Transferase. *Biochemistry* **2007**, *46*, 6353–6363.
- (31) Klähn, M.; Braun-Sand, S.; Rosta, E.; Warshel, A. On Possible Pitfalls in *ab Initio* Quantum Mechanics/Molecular Mechanics Minimization Approaches for Studies of Enzymatic Reactions. *J. Phys. Chem. B* **2005**, *109*, 15645–15650.
- (32) Zhang, Y.; Kua, J.; McCammon, J. A. Influence of Structural Fluctuation on Enzyme Reaction Energy Barriers in Combined Quantum Mechanical/Molecular Mechanical Studies. *J. Phys. Chem. B* **2003**, *107*, 4459–4463.
- (33) Piomelli, D. The Molecular Logic of Endocannabinoid Signaling. *Nat. Rev. Neurosci.* **2003**, *4*, 873–884.
- (34) Lodola, A.; Mor, M.; Hermann, J. C.; Tarzia, G.; Piomelli, D.; Mulholland, A. J. QM/MM Modelling of Oleamide Hydrolysis in Fatty Acid Amide Hydrolase (FAAH) Reveals a New Mechanism of Nucleophile Activation. *Chem. Commun.* **2005**, 439, 9–4401.
- (35) Tubert-Brohman, I.; Acevedo, O.; Jorgensen, W. L. Elucidation of Hydrolysis Mechanisms for Fatty Acid Amide Hydrolase and Its Lys142Ala Variant via QM/MM Simulations. *J. Am. Chem. Soc.* **2006**, *128*, 16904–16913.
- (36) McKinney, M. K.; Cravatt, B. F. Evidence for Distinct Roles in Catalysis for Residues of the Serine-Serine-Lysine Catalytic Triad of Fatty Acid Amide Hydrolase. *J. Biol. Chem.* **2003**, *278*, 37393–37399.
- (37) Stewart, J. J. P. Optimization of Parameters for Semiempirical Methods. 2. Applications. *J. Comput. Chem.* **1989**, *10*, 221–264.
- (38) MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *J. Phys. Chem. A* **1998**, *102*, 3586–3616.
- (39) Lonsdale, R.; Ranaghan, K. E.; Mulholland, A. J. Computational Enzymology. *Chem. Commun.* **2010**, 2354–2372.
- (40) Hermann, J. C.; Ridder, L.; Mulholland, A. J.; Höltje, H. D. Identification of Glu166 as the General Base in the Acylation Reaction of Class A β -Lactamases through QM/MM Modeling. *J. Am. Chem. Soc.* **2003**, *125*, 9590–9591.
- (41) Ridder, L.; Mulholland, A. J.; Vervoort, J.; Rietjens, I. M. C. M. Correlation of Calculated Activation Energies with Experimental Rate Constants for an Enzyme Catalyzed Aromatic Hydroxylation. *J. Am. Chem. Soc.* **1998**, *120*, 7641–7642.
- (42) Wold, S.; Esbensen, K.; Geladi, P. Principal Component Analysis. *Chemom. Intell. Lab. Syst.* **1987**, *2*, 37–52.
- (43) Wold, S.; Sjostrom, M.; Eriksson, L. PLS-Regression: A Basic Tool of Chemometrics. *Chemom. Intell. Lab. Syst.* **2001**, *58*, 109–130.
- (44) Wold, S.; Ruhe, A.; Wold, H.; Dunn, W. J. The Collinearity Problem in Linear Regression. The Partial Least Squares (PLS) Approach to Generalized Inverses. *SIAM J. Sci. Stat. Comput.* **1984**, *5*, 735–743.
- (45) Draper, N. R.; Smith, H. *Applied Regression Analysis*, 2nd ed.; Wiley: New York, 1980.
- (46) Eriksson, L.; Johansson, E.; Kettaneh-Wold, N.; Trygg, J.; Wikstrom, C.; Wold, S. *Multi- and MegaVariate Data Analysis—Basic Principles and Applications*, 2nd ed.; Umetrics AB: Umeå, Sweden, 2006.
- (47) Mor, M.; Rivara, S.; Lodola, A.; Plazzi, P. V.; Tarzia, G.; Duranti, A.; Tontini, A.; Piersanti, G.; Kathuria, S.; Piomelli, D. Cyclohexylcarbamate Acid 3'- or 4'-Substituted Biphenyl-3-yl Esters as Fatty Acid Amide Hydrolase Inhibitors: Synthesis, Quantitative Structure–Activity Relationships, and Molecular Modeling Studies. *J. Med. Chem.* **2004**, *47*, 4998–5008.
- (48) Mor, M.; Lodola, A.; Rivara, S.; Vacondio, F.; Duranti, A.; Tontini, A.; Sanchini, S.; Piersanti, G.; Clapper, J. R.; King, A. R.; Tarzia, G.; Piomelli, D. Synthesis and Quantitative Structure–Activity Relationship of Fatty Acid Amide Hydrolase Inhibitors: Modulation at the N-Portion of Biphenyl-3-yl Alkylcarbamates. *J. Med. Chem.* **2008**, *51*, 3487–3498.
- (49) Valitutti, G.; Duranti, A.; Lodola, A.; Mor, M.; Piersanti, G.; Piomelli, D.; Rivara, S.; Tontini, A.; Tarzia, G.; Traldi, P. Correlation between Energetics of Collisionally Activated Decompositions, Interaction Energy and Biological Potency of Carbamate FAAH Inhibitors. *J. Mass Spectrom.* **2007**, *42*, 1624–1627.
- (50) Tantanak, D.; Limtrakul, J.; Gleeson, M. P. Probing the Structural and Electronic Factors Affecting the Adsorption and Reactivity of Alkenes in Acidic Zeolites Using DFT Calculations and Multivariate Statistical Methods. *J. Chem. Inf. Model.* **2005**, *45*, 1303–1312.
- (51) Gleeson, D. Application of QM Simulations and Multivariate Analysis in the Study of Alkene Reactivity in the Zeolite H-ZSM5. *J. Chemom.* **2008**, *22*, 372–377.
- (52) Fey, N.; Guy Orpen, A.; Harvey, J. N. Building Ligand Knowledge Bases for Organometallic Chemistry: Computational Description of Phosphorus(III)-Donor Ligands and the Metal–Phosphorus Bond. *Coord. Chem. Rev.* **2009**, *253*, 704–722.
- (53) Chatfield, C.; Collins, A. J. *Introduction to Multivariate Analysis*; Chapman and Hall: London, 1980.
- (54) Box, G. E. P.; Hunter, W. G.; Hunter, J. S. *Statistics for Experimenters*; Wiley: New York, 1978.
- (55) Cramer, R. D., III.; Bunce, J. D.; Patterson, D. E. Crossvalidation, Bootstrapping, and Partial Least Squares Compared with Multiple Regression in Conventional QSAR Studies. *Quant. Struct.-Act. Relat.* **1988**, *7*, 18–25.
- (56) *SIMCA-P+*, version 11.0; Umetrics AB: Umeå, Sweden, 2005.
- (57) Bracey, M. H.; Hanson, M. A.; Masuda, K. R.; Stevens, R. C.; Cravatt, B. F. Structural Adaptation in a Membrane Enzyme That Terminates Endocannabinoid Signaling. *Science* **2002**, *298*, 1793–1796.
- (58) Lodola, A.; Mor, M.; Sirirak, J.; Mulholland, A. J. Insights into the Mechanism and Inhibition of Fatty Acid Amide Hydrolase from Quantum Mechanics/Molecular Mechanics (QM/MM) Modeling. *Biochem. Soc. Trans.* **2009**, *37*, 363–367.

- (59) Acevedo, O.; Jorgensen, W. L. Advances in Quantum and Molecular Mechanical (QM/MM) Simulations for Organic and Enzymatic Reactions. *Acc. Chem. Res.* **2010**, *43*, 142–151.
- (60) Patricelli, M. P.; Cravatt, B. F. Clarifying the Catalytic Roles of Conserved Residues in the Amidase Signature Family. *J. Biol. Chem.* **2000**, *275*, 19177–18184.
- (61) Lodola, A.; Mor, M.; Rivara, S.; Christov, C.; Tarzia, G.; Piomelli, D.; Mulholland, A. J. Identification of Productive Inhibitor Binding Orientation in Fatty Acid Amide Hydrolase (FAAH) by QM/MM Mechanistic Modelling. *Chem. Commun.* **2008**, 214–216.
- (62) McKinney, M. K.; Cravatt, B. F. Structure and Function of Fatty Acid Amide Hydrolase. *Annu. Rev. Biochem.* **2005**, *74*, 411–432.
- (63) Hotelling's ellipsoids are equidensity contours resulting from multivariate normal distributions. They describe the area in which an observation can be expected to fall with a certain probability (e.g., 95%).
- (64) Mansson, R. A.; Welsh, A. H.; Fey, N.; Orpen, A. G. Statistical Modeling of a Ligand Knowledge Base. *J. Chem. Inf. Model.* **2006**, *46*, 2591–2600.

CT100264J

Binding Energy Distribution Analysis Method (BEDAM) for Estimation of Protein–Ligand Binding Affinities

Emilio Gallicchio,* Mauro Lapelosa, and Ronald M. Levy

BioMaPS Institute for Quantitative Biology and Department of Chemistry and Chemical Biology, Rutgers the State University of New Jersey, Piscataway, New Jersey 08854

Received June 2, 2010

Abstract: The binding energy distribution analysis method (BEDAM) for the computation of receptor–ligand standard binding free energies with implicit solvation is presented. The method is based on a well-established statistical mechanics theory of molecular association. It is shown that, in the context of implicit solvation, the theory is homologous to the test particle method of solvation thermodynamics with the solute–solvent potential represented by the effective binding energy of the protein–ligand complex. Accordingly, in BEDAM the binding constant is computed by means of a weighted integral of the probability distribution of the binding energy obtained in the canonical ensemble in which the ligand is positioned in the binding site but the receptor and the ligand interact only with the solvent continuum. It is shown that the binding energy distribution encodes all of the physical effects of binding. The balance between binding enthalpy and entropy is seen in our formalism as a balance between favorable and unfavorable binding modes which are coupled through the normalization of the binding energy distribution function. An efficient computational protocol for the binding energy distribution based on the AGBNP2 implicit solvent model, parallel Hamiltonian replica exchange sampling, and histogram reweighting is developed. Applications of the method to a set of known binders and nonbinders of the L99A and L99A/M102Q mutants of T4 lysozyme receptor are illustrated. The method is able to discriminate without error binders from nonbinders, and the computed standard binding free energies of the binders are found to be in good agreement with experimental measurements. Analysis of the results reveals that the binding affinities of these systems reflect the contributions from multiple conformations spanning a wide range of binding energies.

1. Introduction

Molecular recognition is an essential component for virtually all biological processes. In particular, medicinal compounds mainly act by binding to enzymes and signaling proteins, thereby altering their activity. One main aim of drug discovery enterprises is to identify compounds with specific and strong affinity to their target receptors. There is a great interest therefore in the development of computer models capable of predicting accurately the strength of protein–ligand association.¹ Structure-based drug discovery models seek to predict receptor–ligand binding free energies from the

known or presumed structure of the corresponding complex.² Docking and empirical scoring approaches^{3,4} are useful in virtual screening applications^{5,6} but are generally considered not suitable for quantitative binding free energy estimation.

Physical free energy models for binding,⁷ which are built upon realistic representations of molecular interactions and atomic motion, have the potential to achieve sufficient detail and accuracy to address finer aspects of drug development such as ligand optimization and drug specificity, toxicity, and resistance. The computational prediction of protein–ligand binding free energies using these methods remains, however, very difficult due to inaccuracies of the potential models and limitations of conformational sampling as well as to model

* Corresponding author e-mail: emilio@biomaps.rutgers.edu.

uncertainties related to solution conditions and protonation and tautomeric state assignments.⁸ Ongoing development efforts continue to improve the accuracy and usability of free energy models to widen their applicability in drug discovery.

Thermodynamically, the strength of the association between a ligand molecule and its target receptor is measured by the standard binding free energy.⁹ The statistical mechanics theory of molecular association equilibria¹⁰ is nowadays well understood and widely accepted. It provides a prescription to compute standard binding free energies from first principles. Various implementations of this theory exist, some of which, such as free energy perturbation methods,^{11–13} are suitable for estimating relative binding free energies between pairs of similar compounds. A number of methods have been proposed for computing absolute, rather than relative, standard binding free energies. End point approaches compute the free energy of binding by computing the difference between the free energies of the unbound and bound states of the protein–ligand complex. An example of this class of method is the mining minima method,¹⁴ which attempts to exhaustively enumerate and analyze conformations of the ligand and of the complex in terms of their enthalpy and entropic components.¹⁵ Similar in spirit are MM-PBSA/GBSA methods,^{16,17} where enthalpic changes are computed from the analysis of molecular dynamics trajectories. Free energy methods based on the double decoupling¹⁸ and potential of mean force^{19,20} formalisms compute absolute binding free energies by evaluating, with molecular dynamics sampling, free energy estimators along suitable thermodynamic paths connecting the unbound and bound states.²¹ Methods based on the latter involve physically moving the ligand in or out of the receptor, whereas double decoupling methods^{18,22,23} employ alchemical computational techniques to essentially decouple the ligand from the solution environment and make it appear in the receptor site.

In this paper we present a novel approach to binding free energy estimation and analysis we call the binding energy distribution analysis method (BEDAM). One motivation for this work has been our interest in evaluating the performance of implicit solvent modeling in alchemical decoupling strategies which have been traditionally applied in the context of explicit solvation. As part of this work we developed a formalism for the standard free energy of binding based on probability distributions of the binding energy of receptor–ligand complex conformations. We show that this formalism is useful both as an analytical tool to gain insights in the statistical thermodynamics of the binding process as well as for forming the framework for an efficient binding free energy computational algorithm based on parallel Hamiltonian replica exchange conformational sampling and reweighting techniques.

Implicit solvent models,²⁴ which are widely used for protein structure prediction^{25,26} and folding^{27–29} and small molecule hydration,^{30,31} have also been employed in protein–ligand binding studies: for docking and scoring,^{32–36} for linear interaction energy modeling,^{37,38} and for MM-PBSA/GBSA applications as mentioned above as well as for free energy perturbation calculations.³⁹ We developed the analytical generalized Born plus non-polar (AGBNP) implicit

solvent model,⁴⁰ which introduced a number of key innovations with respect to the treatment of electrostatic and nonpolar hydration effects. Recent developments⁴¹ introduced treatment of short-range hydration interactions and improved geometric modeling to achieve a better balance between intramolecular interactions and hydration forces. Because of the parameter-free treatment of geometric estimators (Born radii and atomic surface areas), AGBNP is not only applicable to macromolecules but also to a large variety of drug-like compounds and functional groups. AGBNP includes a model for solute–solvent van der Waals dispersion interactions which is particularly suitable for describing association equilibria⁴² in part because, in contrast to conventional surface area models, it is capable of describing the residual ligand–solvent van der Waals interaction energy in the associated state.⁴³ Together with the availability of analytical gradients and other implementation features, such as multithreading parallelization, these characteristics make AGBNP particularly suitable for molecular dynamics-based modeling of protein–ligand binding.

In the context of this work the “binding energy” of a single conformation of the receptor–ligand complex is defined as the free energy gain or cost of bringing the receptor and ligand from infinite separation in solution to their relative position and orientation in the complex without changing their internal coordinates. In this process the solvent degrees of freedom are averaged and their role is included in the binding energy in terms of the solvent potential of mean force.⁴⁴ Although, in principle, this definition does not depend on whether the solvation treatment is explicit or implicit, in this work we model the solvent potential of mean force by means of an implicit solvent function. This choice is motivated not only by CPU performance but also and much more importantly by the ability to obtain distributions of binding energies over tens of thousands of conformations of the complex, which, as shown below, can be directly employed to estimate the binding free energy. An equivalent calculation with explicit solvation would otherwise require a costly potential of mean force evaluation for each conformation of the complex.

The implicit solvent treatment also allows us to employ the binding energy as a biasing potential on which we build an efficient free energy calculation scheme based on a parallel replica exchange⁴⁵ conformational sampling algorithm and histogram reweighting.⁴⁶ The benefits of replica exchange sampling and multistate reweighting techniques⁴⁷ for free energy estimation has been documented in a variety of contexts^{47–50} including protein–ligand binding free energy estimation.^{51,52} In this work we use this strategy to compute binding energy distributions over a wide range of binding energies and to properly sample the variety of ligand poses that contribute to binding in the system studied here.⁵³

The application of the BEDAM methodology is illustrated on a series of complexes of mutant forms of T4 lysozyme.^{54,55} The small size of the ligands and the relative simplicity of the binding sites, together with the availability of high-quality structural and thermodynamic data,^{55,56} make these systems particularly well suited for validating computational models of protein–ligand binding.⁵⁷ Extensive binding free energy

calculations with explicit solvent have been conducted,^{58,59} which confirmed the applicability (as well as some of the challenges²³) of molecular mechanics modeling aimed at the estimation of binding free energies for this system.

2. Theory and Methods

2.1. Standard Free Energy of Binding from Binding Energy Distributions. We start from the expression of the binding constant, K_{AB} , for the binding of ligand B to receptor A from eq 38 in Gilson et al.¹⁸ relevant for the implicit solvation treatment of the water environment

$$K_{AB} = \frac{C^\circ Z_{AB}}{8\pi^2 Z_A Z_B} \quad (1)$$

where, using the notation in Gilson et al.,¹⁸ C° is the inverse of the standard volume $V^\circ = 1668 \text{ \AA}^3$

$$Z_{AB} = \int d\zeta_B J(\zeta_B) I(\zeta_B) dx_B dx_A e^{-\beta[U(\zeta_B, x_B, x_A) + W(\zeta_B, x_B, x_A)]} \quad (2)$$

is the configurational partition function of the AB complex, and

$$Z_B = \int dx_B e^{-\beta[U(x_B) + W(x_B)]} \quad (3)$$

and

$$Z_A = \int dx_A e^{-\beta[U(x_A) + W(x_A)]} \quad (4)$$

are, respectively, the configurational partition functions of the ligand B and the receptor A in solution. The degrees of freedom for Z_{AB} are the six external coordinates of the ligand (position and orientation) relative to the receptor⁶⁰ which are collectively represented by the variable ζ_B and the internal coordinates of the ligand and receptor which are represented by the variables x_B and x_A , respectively. The configurational partition functions of the ligand and receptor, Z_B and Z_A , extend over the internal degrees of freedom of each binding partner.

In eqs 2–4 β is $1/k_B T$, where k_B is the Boltzmann constant and T is the absolute temperature and U is the potential energy function describing direct covalent and noncovalent intramolecular interactions as well as, for the complex, intermolecular noncovalent interactions between the ligand and the receptor. The function W represents the solvent potential of mean force,⁴⁴ which describes solvent-mediated interactions. In eq 2, $J(\zeta_B)$ represents the Jacobian corresponding to the external coordinates of the ligand relative to the receptor and $I(\zeta_B)$ is an indicator function which defines the complexed state of the system (i.e., $I(\zeta_B) = 1$ within the binding site and $I(\zeta_B) = 0$ outside). As discussed,¹⁸ $I(\zeta_B)$ can be also equivalently defined in terms of a continuous function which interpolates from values near 1 within the binding site region to values near 0 outside, which is the approach we adopt in this work. The expression for K_{AB} given here omits symmetry numbers corrections,⁵³ and consequently, the integrations in the configurational partition functions given here are meant to extend explicitly over all symmetrically equivalent conformations of the ligand.

By multiplying and dividing eq 1 by the quantity

$$V_{\text{site}} = \frac{1}{8\pi^2} \int d\zeta_B J(\zeta_B) I(\zeta_B) \quad (5)$$

which represents the effective volume of the binding site, it is straightforward to show that K_{AB} can be equivalently expressed as¹⁸

$$K_{AB} = C^\circ V_{\text{site}} \langle e^{-\beta u} \rangle_0 \quad (6)$$

with

$$\langle \exp(-\beta u) \rangle_0 = \int d\zeta_B dx_B dx_A \rho_0(\zeta_B, x_B, x_A) e^{-\beta u(\zeta_B, x_B, x_A)} \quad (7)$$

where

$$u(\zeta_B, x_B, x_A) = [U(\zeta_B, x_B, x_A) - U(x_B) - U(x_A)] + [W(\zeta_B, x_B, x_A) - W(x_B) - W(x_A)] \quad (8)$$

is the effective binding energy for a given conformation of the ligand–receptor complex and

$$\rho_0(\zeta_B, x_B, x_A) = \frac{J(\zeta_B) I(\zeta_B) e^{-\beta[U(x_B) + U(x_A)]} e^{-\beta[W(x_B) + W(x_A)]}}{\int d\zeta_B dx_B dx_A J(\zeta_B) I(\zeta_B) e^{-\beta[U(x_B) + U(x_A)]} e^{-\beta[W(x_B) + W(x_A)]}} \quad (9)$$

is the normalized probability distribution of the ensemble of conformations of the complex in the absence of ligand–receptor interactions, including solvent-mediated interactions described by the solvent potential of mean force W .

On the basis of eq 6 the standard binding free energy $\Delta F_{AB}^\circ = -k_B T \log K_{AB}$ can be written as

$$\Delta F_{AB}^\circ = -kT \log C^\circ V_{\text{site}} + \Delta F_{AB} \quad (10)$$

where the first term is interpreted as the entropic work corresponding to the process of transferring the ligand from a solution of concentration C° to the binding site region of the complex. This term depends only on the definition of the standard state and the definition of the complex macrostate (according to the indicator function $I(\zeta_B)$) and does not depend on any specific energetic property of the receptor and the ligand. The second free energy term in eq 10 defined as

$$\Delta F_{AB} = -kT \log \langle e^{-\beta u} \rangle_0 \quad (11)$$

represents the work for turning on interactions between the ligand and the receptor while the ligand is sequestered within the binding site region. More precisely, ΔF_{AB} corresponds to the difference in free energy between a fictitious state (henceforth referred as the “solvated reference” state) in which the ligand and the receptor do not see each other (even though the ligand is confined within the binding site) and interact solely with the solvent continuum and a “bound” state in which the receptor and the ligand see each other in terms of direct electrostatic and van der Waals interactions as well as in terms of mutual desolvation effects due to the displacement of the

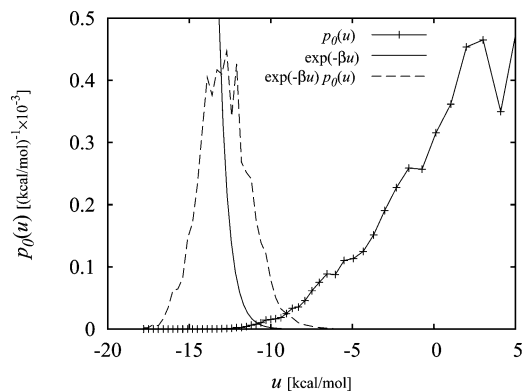


Figure 1. Calculated binding energy distribution $p_0(u)$ for the complex between benzene and the L99A mutant of T4 lysozyme. The curves to the left correspond to the $\exp(-\beta u)$ and $\exp(-\beta u)p_0(u)$ functions (rescaled to fit within the plotting area). The integral of the latter is proportional to the binding constant (eq 12).

solvent continuum from each other's environments. Note that unlike the binding site volume term in eq 10, ΔF_{AB} is independent of the definition of the standard state. As expressed by eq 10, the combination of the two processes of transferring the ligand in the binding site region and turning on receptor–ligand interactions is thermodynamically equivalent to the binding of the ligand to the receptor.

A useful representation for the binding constant (or equivalently for the standard binding free energy ΔF_{AB}°) is obtained by writing the average $\langle \exp(-\beta u) \rangle_0$ in eq 6 in terms of a probability distribution density of the binding energy

$$K_{AB} = e^{-\beta \Delta F_{AB}} = C^\circ V_{\text{site}} \int du p_0(u) e^{-\beta u} \quad (12)$$

where $p_0(u)$ formally defined as

$$p_0(u) = \langle \delta[u(\zeta_B, x_B, x_A) - u] \rangle_0 \quad (13)$$

is the probability distribution for the binding energy in the solvated reference state (the same conformational ensemble specified above by eq 9).

The calculated binding energy probability distribution functions $p_0(u)$ for one of the ligands of mutant T4 lysozyme discussed below are shown in Figure 1. As illustrated in this figure, $p_0(u)$ is largest for large positive values of u with a low-probability tail extending to negative values of u . This is expected since in the absence of receptor–ligand interactions the ligand is more likely to sample conformations with unfavorable clashes between receptor and ligand atoms rather than conformations with favorable interactions with the receptor. The values of u in the extreme negative binding energy range correspond to low-energy conformations of protein–ligand complexes such as those provided by X-ray crystallography and ligand docking. As illustrated in Figure 1, while $p_0(u)$ increases with increasing u , the $\exp(-\beta u)$ function decreases rapidly in the same direction. In order for the integral of the product of these two functions to be finite, it is necessary for $p_0(u)$ to decrease faster than exponentially for $u \rightarrow -\infty$. As shown below, the computed $p_0(u)$ functions in this work satisfy this asymptotic limit. In

addition, the assumed normalization property imposes the requirement that $p_0(u)$ decays faster than $1/u$ for $u \rightarrow \infty$.

The integral in eq 12 is dominated by the tail of the distribution at favorable values of u where $\exp(-\beta u)$ is large and $p_0(u)$ is not negligible (see Figure 1). This however should not be taken to imply that the bulk of conformations that occur at unfavorable values of the binding energy have no effect on the resulting binding free energy. Because $p_0(u)$ is normalized, conformations at unfavorable binding energies oppose binding by increasing the magnitude of the distribution at unfavorable binding energies at the expense of the magnitude of the favorable binding energy tail of the distribution. The specific behavior of $p_0(u)$ at large u 's, however, is not significant because that region of the (properly normalized) distribution makes a negligible contribution to the integral of eq 12. The latter is an important feature for the computational implementation of the theory because in practice, due to the sparseness of the collected samples at large binding energies, it is not feasible to estimate precisely the shape of the distribution at large values of u . Knowledge of the cumulative probability $P_0(u > u_{\text{max}})$ of observing any unfavorable binding energy larger than an appropriate large value u_{max} is sufficient to obtain accurate estimates of the binding free energies (see the section on Details of Computer Simulations below for details on the binning procedure we employed).

It should be noted that the formalism described above is homologous to the potential distribution theorem (PDT)^{61,62} of which the particle insertion method of solvation thermodynamics⁶³ is a particular realization.⁶⁴ In particle insertion the standard chemical potential of the solute, μ , is written in terms of the probability distribution $p_0(v)$ of solute–solvent interaction energies, v , corresponding to the ensemble in which the solute is not interacting with the solvent

$$e^{-\beta \mu} = \int dv p_0(v) e^{-\beta v} \quad (14)$$

This expression, except for the term $C^\circ V_{\text{site}}$, is equivalent to eqs 6 and 12 with the solute–solvent interaction energy v replaced by the protein–ligand binding energy u . It follows that the formalism described above for the binding free energy can be regarded as a ligand insertion theory for protein–ligand binding, where the protein atoms and the solvent continuum play the same role as the solvent molecules in particle insertion.

A known result of particle insertion theory is a relationship between $p_0(v)$, the probability distribution of solute–solvent interaction energies in the absence of solute–solvent interactions, and $p_1(v)$, the corresponding probability distribution in the presence of solute–solvent interactions.⁶⁵ In the present notation we have

$$p_1(v) = e^{\beta \mu} e^{-\beta v} p_0(v) \quad (15)$$

where μ is the chemical potential. The corresponding expression linking $p_0(u)$, the probability distribution of ligand–protein binding energies for the solvated reference state, and $p_1(u)$, the probability distribution for the bound state is

$$p_1(u) = e^{\beta\Delta F_{AB}} e^{-\beta u} p_0(u) \quad (16)$$

where ΔF_{AB} , defined by eq 11, is the interaction-dependent component of the standard binding free energy. It follows that $p_1(u)$ is proportional to the integrand in eq 12 for the binding free energy. Note however that this does not imply that the binding free energy can be computed by integration of $p_1(u)$, as obtained, for example, from a conventional simulation of the complex in the presence of ligand–receptor interactions. The integral of the normalized probability distribution $p_1(u)$, which is by definition unitary, does not contain any information about the binding free energy. As expressed by eq 16, the proportionality constant between $p_1(u)$ and the integrand of eq 12 is related to the binding free energy, which is exactly the quantity we are seeking to compute. As discussed below, $p_1(u)$ is nevertheless a useful quantity for analysis of the relative contributions to the binding free energy of macrostates of the complex.

2.2. Binding Affinity Density. According to eq 12 the binding constant can be expressed in terms of an integral over the function

$$k(u) = C^{\circ} V_{\text{site}} e^{-\beta u} p_0(u) \quad (17)$$

which can be interpreted as a measure of the contribution of the conformations of the complex with binding energy u to the binding constant. We thus call the function $k(u)$ the binding affinity density.

Comparison of eqs 16 and 17 leads to the conclusion that the binding affinity density $k(u)$ is proportional to $p_1(u)$, the binding energy probability distribution in the ligand-bound state. (The critical distinction between the two is that the integral of the latter is equal to 1 whereas the integral of the binding affinity density is equal to the binding constant.) It thus follows that the relative contributions to the binding constant of two complex macrostates one with binding energy u_1 and another with binding energy u_2 is simply given by their relative populations in the ligand-bound state when the interactions between the ligand and the receptor are fully turned on.

Figure 10 shows the calculated binding affinity densities for some of the complexes studied in this work. The densities of higher magnitude and larger subtended area correspond to more tightly bound complexes. The corresponding $p_1(u)$ distributions, since by definition they all subtend the same surface area, have the same shape but with much smaller differences in magnitude across the various ligands.

2.3. Conformational Decomposition. Given a set of macrostates $i = 1, \dots, n$ of the complex, corresponding, for example, to different ligand poses in the receptor site, we consider the joint probability distribution $p_0(u, i)$, expressing the probability of observing the binding energy u while the complex is in macrostate i . Assuming that the set of macrostates collectively covers all possible conformations of the complex (which is always possible by including a catch-all macrostate), we can express $p_0(u)$ as a marginal of $p_0(u, i)$

$$p_0(u) = \sum_i p_0(u, i) = \sum_i P_0(i) p_0(ul_i) \quad (18)$$

where we introduced the conditional distribution $p_0(ul_i)$ and the population $P_0(i)$ of macrostate i in the solvated reference state and used the relationship $p_0(u, i) = P_0(i) p_0(ul_i)$ between the joint and conditional distributions. By inserting eq 18 into eq 17, we have

$$k(u) = \sum_i P_0(i) k_i(u) \quad (19)$$

where

$$k_i(u) = C^{\circ} V_{\text{site}} p_0(ul_i) e^{-\beta u} \quad (20)$$

represents the binding affinity density for macrostate i . In analogy with eqs 10 and 17 we define a macrostate-specific binding constant

$$K_{AB}(i) = e^{-\beta\Delta F_{AB}(i)} = \int du k_i(u) = C^{\circ} V_{\text{site}} \langle e^{-\beta u} \rangle_{0,i} \quad (21)$$

where $\langle \dots \rangle_{0,i}$ represents an ensemble average in the solvated reference state limited to macrostate i . The macrostate-specific binding constant $K_{AB}(i)$ represents therefore the binding constant that would be measured if the conformations of the complex are limited to macrostate i . From eqs 21 and 19 the sum of the macrostate-specific binding constants weighted by the macrostate populations $P_0(i)$ is the total binding constant

$$K_{AB} = \sum_i P_0(i) K_{AB}(i) \quad (22)$$

The ratio $P_0(i) K_{AB}(i) / K_{AB}$ (reported in Figure 11 for the systems studied here) measures the relative contribution of macrostate i to the overall binding constant. It is straightforward to show from eqs 21 and 16 that

$$\frac{P_0(i) K_{AB}(i)}{K_{AB}} = P_1(i) \quad (23)$$

where

$$P_1(i) = \int du p_1(u, i) \quad (24)$$

is the population of macrostate i in the bound state. In other words, this analysis shows that the relative contribution of macrostate i to the binding constant is equal to the physical population of that macrostate of the complex.

Similar to previous analysis,⁶⁶ eq 22 expresses the fact that each conformational macrostate contributes to the total binding constant proportionally to its macrostate-specific binding constant $K_{AB}(i)$ weighted by the population of the macrostate in the solvated reference state measured by $P_0(i)$. Similar decompositions have also been previously employed.⁵³ In this work (see Figure 11) we analyze our results in terms of the relative contributions of each macrostate to the total binding constant using eq 23, and we also report the macrostate-specific binding free energies, $\Delta F_{AB}^{\circ}(i)$ from eq 21, for the major macrostates of the system defined as described below.

2.4. Numerical Considerations. The computation of V_{site} from eq 5 is straightforward as it involves integration over only the six degrees of freedom ζ_B , which completely specify

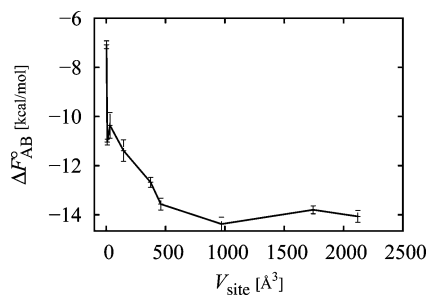


Figure 2. Standard binding free energy between phenol and the L99A/M102Q receptor as a function of the binding site volume. These calculations employed a simple distance-dependent dielectric model of solvation.

the positioning and orientation of the ligand relative to the receptor.⁶⁰ For the calculations carried out in this work we adopt an analytical expression for V_{site} corresponding to the particular choice for the indicator function $I(\zeta_{\text{B}})$ (see below).

As conjectured by Gilson et al.,¹⁸ the value of the standard binding free energy estimated from eq 1 depends only weakly on the specific definition of the $I(\zeta_{\text{B}})$ indicator function as long as this includes all of the important regions of the binding site volume and that the binding is sufficiently strong and specific. We confirmed numerically this conjecture for one of the T4 lysozyme complexes systems studied in this work by performing binding free energy calculations as a function of binding site volume (Figure 2). The results indeed show that the binding free energy reaches a plateau at a binding site volume of approximately 450 \AA^3 and that further increases of the binding site region do not significantly alter the results. This is a consequence of the fact that the binding site volume beyond the natural dimensions of the pocket (which can be estimated as approximately 450 \AA^3 based on Figure 2) only allows additional poses of the ligand that clash with receptor atoms and that, therefore, contribute only repulsive ($u > 0$) binding energies. It is easy to see that in this regime, as the binding site volume increases, the shape of the binding energy distribution $p_0(u)$ at favorable binding energies ($u < 0$) remains unchanged while its magnitude decreases due to the change of normalization, which is in turn proportional to the increase in binding site volume. It follows that the integral over the binding energy distribution in eq 12, of which the $u < 0$ range is the main contributor, decreases as $1/V_{\text{site}}$ for sufficiently large V_{site} . This dependence is exactly canceled by the $C^\circ V_{\text{site}}$ prefactor in eq 12, thereby leading to the observed constancy of the binding constant with increasing binding site volume (Figure 2).

Increasing the binding site volume further could give the ligand access to alternative binding sites on the protein surface, potentially causing changes to the computed binding constant in ways not addressed by the arguments discussed above. For an in-depth discussion of the relationship between the microscopic definition of the binding constant and macroscopic observables of binding we refer the reader to the study of Mihailescu and Gilson.⁶⁷

Having defined the binding site volume, the problem of computing the standard free energy of binding ΔF_{AB}^0 is reduced to the problem of evaluating ΔF_{AB} with eq 11. It is apparent from eqs 10, 11, and 12 that, given a definition of

the ligand-bound macrostate, $p_0(u)$ encodes all of the information necessary to specify the standard binding free energy of the protein–ligand complex. In principle, the calculation of $p_0(u)$ can be accomplished by brute-force collection of binding energy values from a simulation of the complex in the absence of receptor–ligand interactions (with the exception of the binding-site restraints specified by the indicator function $I(\zeta_{\text{B}})$). However, this strategy would produce large finite sampling errors for the binding free energy through eq 12.⁶⁸ The integral in eq 12 is dominated by the favorable binding energy tail of $p_0(u)$, which is rarely sampled when the ligand is not guided by the interactions with the receptor. Inaccuracies in the tail of the distribution are in turn amplified by the $\exp(-\beta u)$ function, thereby affecting the reliability of the free energy estimate. As discussed below, biased sampling combined with the weighted histogram analysis method (WHAM)⁴⁶ provides a very efficient strategy to compute $p_0(u)$ with high precision on a wide range of binding energies, leading to well-converged estimates for ΔF_{AB} from eq 12. The reliability of this strategy is illustrated, for example, in Figure 1, which shows that $p_0(u)$ evaluated by WHAM is sufficiently well defined over the range of binding energies in which $p_1(u) \propto \exp(-\beta u)p_0(u)$ is non-negligible.⁹

2.5. Binding Energy-Biased Conformational Sampling.

As discussed above, straightforward binning of the binding energy values at the unbound thermodynamic end point of the binding process leads to poor estimation of the favorable binding energy tail of the binding energy distributions, which are important for accurate computation of the binding constant. To address this problem we employ biased simulations which, collectively, are able to uniformly sample a wide range of binding energies. The results of these biased simulations are processed using WHAM to produce the unbiased binding energy distribution $p_0(u)$ that are integrated using eq 12 to yield the binding constant.

The biased potential energy ansatz that we employ is of the form

$$V_\lambda = V_0 + \lambda u \quad (25)$$

where λ is the free energy progress parameter and

$$V_0 = V_0(x_{\text{A}}, x_{\text{B}}) = U(x_{\text{A}}) + W(x_{\text{A}}) + U(x_{\text{B}}) + W(x_{\text{B}}) \quad (26)$$

is the effective potential energy of the complex in the absence of direct and solvent-mediated ligand–receptor interactions and $u = u(\zeta_{\text{B}}, x_{\text{A}}, x_{\text{B}})$ is the binding energy of a given conformation of the complex as defined by eq 8. It is easy to see from eqs 2–4, 8, and 26 that $V_{\lambda=1}$ corresponds to the effective potential energy of the bound complex and $V_{\lambda=0}$ corresponds to the state in which the receptor and ligand are not interacting. Intermediate values of λ trace an alchemical thermodynamic path connecting these two states. Multiple simulations at different values of λ are performed along this path, which collectively sample a wide range of unfavorable, intermediate, and favorable binding energies which can be employed with WHAM to estimate with high precision the binding energy probability distribution at $\lambda =$

0. From eqs 25 and 26 it follows that the biasing potential $w_\lambda = V_\lambda - V_0$ required in the application of the WHAM formula takes the simple form

$$w_\lambda(\zeta_B, x_A, x_B) = \lambda u(\zeta_B, x_A, x_B) \quad (27)$$

that is the biasing potential is proportional to the binding energy itself. With this result, unbiased binding energy distributions are obtained by iterative application of the WHAM formula⁴⁶

$$p_0(u) = \frac{n(u)}{\sum_\lambda n_\lambda f(\lambda) \exp(-\beta \lambda u)} \quad (28)$$

where

$$f(\lambda)^{-1} = \sum_u \exp(-\beta \lambda u) p_0(u) \quad (29)$$

$n(u)$ is the number of samples from all simulations with binding energy within the bin corresponding to the binding energy value u , and n_λ is the total number of samples from the simulation at λ . We confirmed that the WHAM equations as implemented are numerically capable of correctly representing the large dynamic range of probabilities necessary to describe $p_0(u)$ (see, for example, Figure 8). The joint probability $p_0(u, i)$ for the conformational decomposition analysis is similarly obtained by WHAM considering the computed histograms $n(u, i)$ corresponding to conformations of the complex in macrostate i with binding energy u . The macrostate populations $P_0(i)$ are obtained by integration of $p_0(u, i)$ over u .

2.6. Hamiltonian Replica Exchange Sampling. To enhance the sampling efficiency of ligand conformations within the receptor binding site it is useful to couple the simulations at different λ values above using an Hamiltonian parallel replica exchange scheme (HREM). In this scheme pairs of simulation replicas periodically attempt to exchange λ values through Monte Carlo (MC) λ -swapping moves. MC attempts are accepted with probability

$$\Pi_{12} = \min(1, e^{-\beta \Delta_{12}}) \quad (30)$$

with

$$\Delta_{12} = -(\lambda_2 - \lambda_1)(u_2 - u_1) \quad (31)$$

where u_2 and u_1 are the binding energies of the pair of replicas and λ_2 and λ_1 are their respective λ values before the attempted exchange.

The benefit of the HREM scheme in λ space is illustrated in Figure 3, which shows the computed binding free energies, using coupled and uncoupled simulations, of phenol bound to the L99A/M102Q mutant of T4 lysozyme as a function of simulation time. (These benchmark calculations conducted with a simple distance-dependent dielectric force field significantly overestimate the magnitude of the binding free energy of phenol, but they nevertheless illustrate the advantages of HREM for this application.)

One set of simulations was started from a conformation similar to the crystal structure, and another set was started

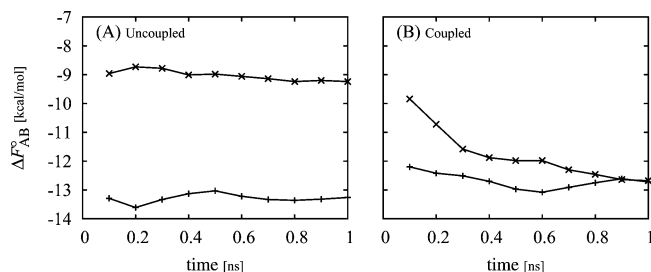


Figure 3. Standard free energy of binding of phenol to the L99A T4 lysozyme receptor with two different starting conditions as a function of simulation time from uncoupled umbrella sampling simulations (A) and from a coupled parallel Hamiltonian replica exchange simulation (B). Plus symbols (+) correspond to simulations started from the crystallographic conformation (PDB id 1LI2) and crosses correspond (x) to simulations started from a noncrystallographic conformation in which phenol is not hydrogen bonded to Q102. These calculations employed a simple distance-dependent dielectric model of solvation.

from another conformation lacking the hydrogen bond between phenol and Q102 of the receptor which is known to be critical for strong binding. We see from Figure 3A that the uncoupled simulations started from the noncrystallographic conformation yield binding free energies less favorable than uncoupled simulations started from the crystallographic simulation. HREM instead (see Figure 3B) yields binding free energies that converge to the same value regardless of the starting conformation. The reason for this behavior is that in the HREM scheme the ligand is less likely to become trapped in low-energy conformations when the ligand interacts strongly with the receptor at $\lambda \approx 1$. For example, in the uncoupled simulation at $\lambda = 1$, which corresponds to a conventional simulation of the complex, phenol is observed to remain in the starting conformation for nearly the entire duration of the longest simulation. Kinetic trapping at large λ leads to poor convergence as shown in Figure 3A, where the uncoupled simulation started from the noncrystallographic conformation grossly underestimates the magnitude of the binding free energy, whereas the one started from the crystallographic conformations overestimates it by a small amount. In contrast, HREM does not suffer from kinetic trapping to the same extent because λ exchanges allow trapped conformations at large λ to assume smaller values of λ which facilitate transitions between different ligand conformations thanks to the weaker interactions with the receptor. By further random exchanges, these new ligand conformations can then assume again large λ values ultimately yielding more extensive conformational sampling at both small and large values of λ .

2.7. Details of Computer Simulations. The T4 lysozyme protein receptors and their respective ligands are shown in Figures 4 and 5. We considered eight ligands for each receptor (16 total), one-half of which are known binders and one-half are known nonbinders.^{23,55,56} For each receptor, initial structures for the complex of benzene with the L99A mutant of T4 lysozyme and that of phenol bound to the L99A/M102Q mutant were prepared based on the corresponding crystal structures (PDB access codes 3DMX and 1LI2, respectively). The initial structures for all of the other

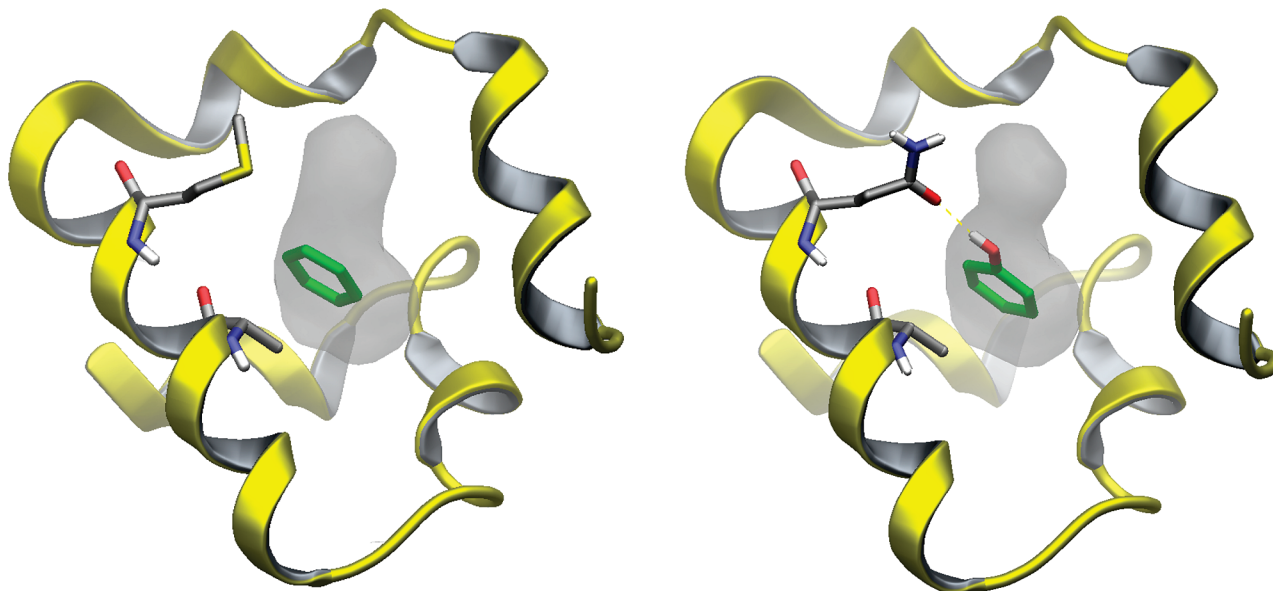


Figure 4. Crystal structures of the benzene-L99A (PDB id 3DMX, left) and phenol-L99A/M102Q (PDB id 1LI2, right) T4 lysozyme complexes. The A99 and M102 residues (Q102 for the L99A/M102Q receptor) are indicated. Residues 73–125 of T4 lysozyme are represented by the ribbon diagram. The ligand is highlighted in green. The surface surrounding the ligand represents the cavity created by the L99A and L99A/M102Q mutations.

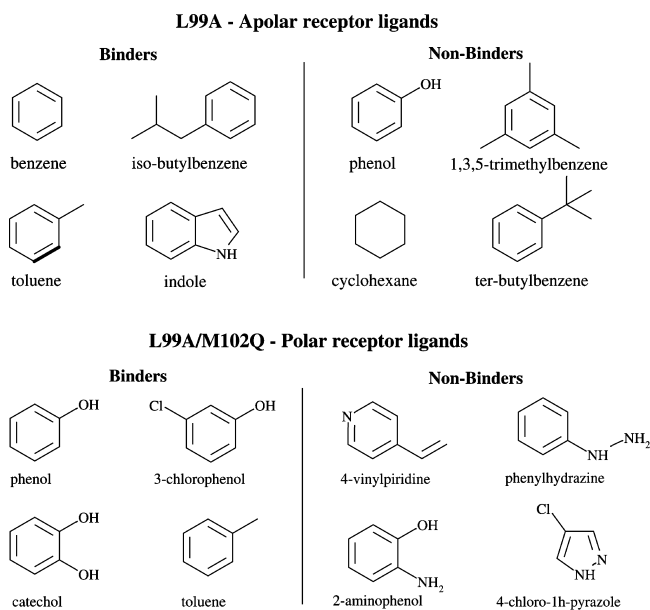


Figure 5. T4 lysozyme ligands investigated in this work.

complexes were prepared by superimposition of each ligand onto the conformations of either benzene or phenol. Hydrogen atoms were added and ionization states assigned assuming neutral pH. The position of C_{α} atoms was restrained near their crystallographic positions with an isotropic quadratic function with force constant $k_f = 0.6 \text{ kcal/mol/\AA}^2$, which allows for approximately a 4 Å range of motion at the simulation temperature. The other backbone atoms and protein side chains were allowed to move freely.

We employ the OPLS-AA⁶⁹ force field with the AGBNP2⁴¹ implicit solvent model. AGBNP2 is a recent evolution of the AGBNP implicit solvent model,⁴⁰ which is based on a parameter-free analytical implementation of the pairwise descreening scheme of the generalized Born model⁷⁰ for the electrostatic component, and a nonpolar hydration

free energy estimator for the nonelectrostatic component. Unlike traditional models based only on the solute surface area, the nonpolar term in AGBNP is the sum of two distinct estimators, one designed to mimic solute–solvent van der Waals dispersion interactions and a second corresponding to the work required for the formation of the solute cavity in water. The AGBNP2 model includes a novel first solvation shell function to improve the balance between solute–solute and solute–solvent interactions based on the results of benchmark tests with explicit solvation. The AGBNP2 model also introduces an analytical solvent-excluded volume model which improves the solute volume description by reducing the effect of spurious high-dielectric interstitial spaces present in conventional van der Waals volume representations.⁴¹

Each complex was energy minimized and thermalized at 310 K. λ -biased replica exchange molecular dynamics simulations were conducted for 2 ns with a 1.5 fs MD time step at 310 K with 12 replicas at $\lambda = 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 0.15, 0.25, 0.5, 0.75, 1,$ and 1.2. The parallel replica exchange calculations took approximately 30 h per complex on 96 processor cores using a custom multithreaded version of the IMPACT program.⁷¹ Bond lengths with hydrogen atoms were constrained using SHAKE. The mass of hydrogen atoms was set to 5 amu. A 12 Å residue-based cutoff was imposed on both direct and generalized Born pair interactions. Soft-core potentials were employed for both Lennard–Jones and Coulomb interactions using a modified distance function of the form $r' = (r^{12} + a^{12})^{1/12}$ with $a = 1 \text{ Å}$. This modified distance function limits the magnitude of nonbonded interactions at short interatomic distances that occur at small λ ; it has a negligible effect on the interaction energies at the interatomic distances normally encountered with full ligand–receptor interactions (for example, at $r = 1.5 \text{ Å}$, a distance typical for the shortest

nonbonded interactions, the modified distance is only 0.06% larger than the actual distance).

Each replica simulates the complex with a biased potential of the form shown in eq 25. Using eqs 26 and 8 it is straightforward to show that eq 25 corresponds to a hybrid potential of the form

$$V_\lambda = U(x_A) + U(x_B) + \lambda U(\zeta_B, x_A, x_B) + \lambda W(\zeta_B, x_A, x_B) + (1 - \lambda)[W(x_A) + W(x_B)] \quad (32)$$

where $U(x_A)$ and $U(x_B)$ represent the intramolecular interactions (Lennard–Jones and Coulomb interactions) of the receptor and the ligand and $U(\zeta_B, x_A, x_B)$ represents their mutual interactions. $W(\zeta_B, x_A, x_B)$, $W(x_A)$, and $W(x_B)$ are, respectively, the AGBNP2 hydration free energies of the complex, the receptor, and the ligand. It is straightforward to implement the nonbonded component of eq 32 by rescaling direct receptor–ligand interactions during the simulation. The implicit solvent components are currently implemented as two separate invocations of the routines for the AGBNP2 energy and gradients, one for the complex and one for the separated receptor and ligand.

Protein–ligand binding energies of each replica were collected every 1 ps during the second half of the simulation. The replica exchange simulations yielded a total of 12 000 binding energy samples for each ligand that were employed to compute an overall histogram $n(u)$ of binding energies. One hundred ten histogram bins were employed with increasing bin spacing for increasing values of u from -30 to 80 kcal/mol. Values of u larger than this maximum were counted toward the last bin. Histograms were processed through the WHAM eq 28 with the biasing function (eq 27) to yield the binding energy distributions $p_0(u)$. These were then integrated according to eq 12 to yield the standard binding free energy values ΔF_{AB}^λ for each ligand. Statistical uncertainties were computed by block bootstrap analysis²³ on the set of computed binding energies using 8 samples.

The binding site indicator function was set as $I(\zeta_B) = \exp[-\beta\omega(r, \cos \theta, \phi)]$, see eq 2, where $\omega(r, \cos \theta, \phi)$ is a product of flat-bottom harmonic potentials acting on the position, expressed in polar coordinates,⁶⁰ of one of the atoms of the aromatic ring of each ligand with respect to the positions of the C_α atoms of residues 88, 102, and 111 of the receptor. The distance restraint potential was centered at 6.4 \AA with a 5 \AA tolerance on either side (allowing unhindered distances from 1.4 to 11.4 \AA). Distances beyond these limits were penalized by means of a quadratic function with a force constant of 3 kcal/mol/\AA^2 . The flat-bottom harmonic restraint potential for the cosine of the angle θ between the reference ligand atom, the C_α atom of residue 88, and the C_α atom of residue 102 was centered at $\cos \theta = 0.85$ with a 0.15 tolerance on either side and a force constant of 100 kcal/mol beyond that. Finally, the restraint potential for the dihedral angle defined by the three atoms above plus the C_α atom of residue 111 was centered at $\phi = 20^\circ$ with a 50° tolerance on either side and a force constant of 0.1 kcal/mol/deg beyond that. The variables corresponding to the orientation of the ligand with respect to the receptor were not restrained; they contribute $8\pi^2$ to the integral in eq 5,

Table 1. Experimental and Calculated Standard Binding Free Energies and Corresponding Ligand Rankings

molecule	$\Delta F^\circ(\text{expt})^{a,b}$	$\Delta F^\circ(\text{calcd})^a$	rank (expt)	rank (calcd)
L99A apolar cavity				
iso-butylbenzene	-6.51^c	-5.21 ± 0.06	1	1
toluene	-5.52^c	-3.80 ± 0.05	2	3
benzene	-5.19^c	-4.01 ± 0.04	3	2
indole	-4.89^c	-3.75 ± 0.02	4	4
tert-butylbenzene	$>-2.7^d$	-2.93 ± 0.03		5
cyclohexane	$>-2.7^d$	-2.21 ± 0.05		6
1,3,5-trimethylbenzene	$>-2.7^d$	-1.68 ± 0.05		7
phenol	$>-2.7^d$	-1.40 ± 0.03		8
L99A/M102Q polar cavity				
3-chlorophenol	-5.51^e	-3.47 ± 0.05	1	3
phenol	-5.23^e	-3.65 ± 0.04	2	2
toluene	-4.93^e	-4.26 ± 0.06	3	1
catechol	-4.16^e	-3.44 ± 0.04	4	4
4-vinylpyridine	$>-2.7^e$	-2.38 ± 0.02		5
4-chloro-1h-pyrazole	$>-2.7^e$	-1.60 ± 0.03		6
2-aminophenol	$>-2.7^e$	-0.70 ± 0.05		7
phenylhydrazine	$>-2.7^e$	2.63 ± 0.05		8

^a In kcal/mol. ^b A lower-limit estimate given for nonbinders.⁵⁶ ^c Reference 56. ^d Reference 55. ^e Reference 23.

thereby canceling out the same quantity in the denominator of eq 5. According to this definition of $I(\zeta_B)$, the volume of the binding site V_{site} (eq 5) was measured to be 469.2 \AA^3 corresponding to a value for $-k_B T \ln C^\circ V_{\text{site}}$ in eq 10 of approximately 0.75 kcal/mol . This value, which is the same for all ligands, is added to the value of the computed ΔF_{AB} for each ligand to yield the standard binding free energies reported in Table 1.

The macrostates of the complex for the conformational decomposition analysis have been defined in terms of the orientation of the ligand with respect to a reference orientation (based typically on the crystallographic structure). The central binding site cavities of the two receptors which contain the aromatic ring of the ligand (Figure 4) are wide and flat, allowing basically only two possible kinds of motion of the ligand: rotation within the plane of the ring and a 180° flip of the plane of the ring. These motions are captured, respectively, by the pitch angle θ_n between the normals to the ring planes of the reference and given conformation of the ligand and the in-plane rotation angle θ_p between the given and reference axes going through a chosen atom of the ring (Figure 6).

Macrostate boundaries are selected from the distribution of samples of (θ_n, θ_p) pairs collected from the HREM replica at $\lambda = 1$. A representative example is given in Figure 7 for phenol bound to the L99A/M102Q receptor. Two macrostates can be identified, one corresponding to the crystallographic pose with $\theta_p = 0^\circ \pm 30^\circ$ and another less populated macrostate with $\theta_p = -60^\circ \pm 30^\circ$. In this case, given the C_2 symmetry of phenol, the θ_n angles near 0° and 180° correspond to the same state. The difference in the number of samples between the left and the right sides of Figure 7 can be used, after taking into account statistical fluctuations, as a measure of convergence of the HREM conformational sampling protocol. For molecules lacking C_2 symmetry, such as 3-chlorophenol, the θ_n angle is used to distinguish conformations with substituents oriented on opposite sides of the ring. For all ligands the definition of the macrostates

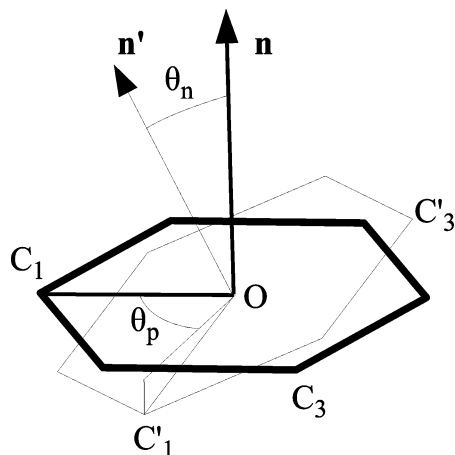


Figure 6. Diagram depicting the definition of the pitch angle θ_n and in-plane rotation angle θ_p used in the conformational decomposition analysis. The hexagon in thick lines represents the aromatic ring of the reference pose, C_1 and C_3 are two atoms of the ring, O is the centroid of the heavy atoms of the ring, and \mathbf{n} is the normal to the plane of the ring (the plane defined by O , C_1 , and C_3). C'_1 , C'_3 , and \mathbf{n}' are the corresponding quantities for the ring of the given pose. θ_n is defined as the angle between \mathbf{n} and \mathbf{n}' , and θ_p is defined as the angle between the OC_1 segment and the projection of the OC'_1 segment onto the plane of the ring of the reference pose.

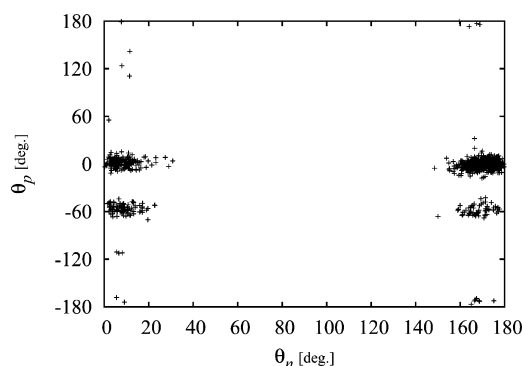


Figure 7. Samples of pitch and in-plane rotational angles pairs (θ_n, θ_p) for phenol bound to the L99A/M102Q T4 lysozyme receptor.

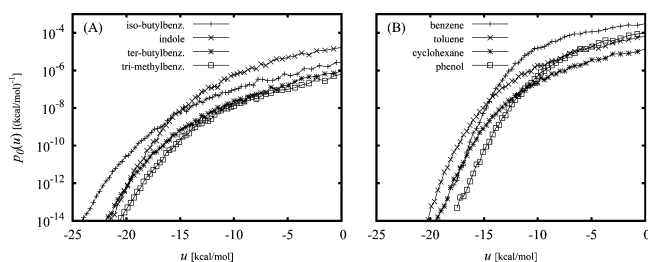


Figure 8. Favorable binding energy tails of the binding energy distributions of the L99A T4 lysozyme complexes.

used a range of 30° on either side of a central value of the in-plane θ_p rotation angle identified similarly as for phenol above. For molecules possessing C_2 symmetry, the ranges of the pitch angle θ_n included the intervals $\theta_n < 30^\circ$ and $\theta_n > 150^\circ$. For other molecules the range for θ_n includes only one of the two intervals depending on the macrostate.

3. Results

The computed binding energy distributions obtained from the BEDAM calculations are shown in Figures 1, 8, and 9. The corresponding standard binding free energies from eq 12 for the L99A and L99A/M102Q mutants of T4 lysozyme are presented in Table 1 for the ligands listed in Figure 5. We see that the ligand rankings based on the computed binding free energies distinguish without errors the binders from the nonbinders as determined experimentally. For example, the model correctly predicts that toluene binds to both the L99A and the L99A/M102Q receptors while phenol binds only to the L99A/M102Q receptor. More subtle trends are also reproduced. Iso-butylbenzene is correctly predicted as the best binder to the L99A receptor, while the binding of the relatively similar *tert*-butylbenzene is correctly predicted to be much weaker. Cyclohexane is correctly predicted as a nonbinder of the L99A receptor, distinguishing it from benzene, which is a binder. The related catechol and 2-aminophenol are correctly differentiated as a binder and nonbinder, respectively, to the L99A/M102Q receptor.

The method correctly reproduced the ranking of the best binder (iso-butylbenzene) and the weakest binder (indole) of the L99A receptor, whereas the rankings of the two intermediate binders, benzene and toluene, are reversed relative to the experiments. The order of the rankings of the binders to the L99A/M102Q receptor are not as accurate relative to the experiments. Toluene is predicted to be the best binder for the L99A/M102Q receptor, whereas 3-chlorophenol is known to be the best binder in this set.

The computed standard binding free energies all underestimate the experimental binding affinities. For the L99A receptor the amount of underestimation is approximately 1.2 kcal/mol for most of the binders. Relative binding free energies are in good agreement with the experiments. Larger variations in accuracy are observed for the L99A/M102Q receptor binders with toluene having the smallest discrepancy (approximately 0.7 kcal/mol), while larger discrepancies are observed for the polar compounds (up to approximately 2 kcal/mol for 3-chlorophenol).

The binding energy distributions provide insights into the binding thermodynamics of these complexes. Figures 8 and 9 show, in logarithmic scale, the details of the low binding energy tails of the computed binding energy distributions for the L99A and L99A/M102Q complexes, respectively. As discussed above, this region of the distributions provides nearly all of the contribution to the binding affinity. It can be clearly seen from these results that the $p_0(u)$ distributions decay with decreasing binding energy faster than exponential (that is faster than linear in the log scale) as required by the theory. The ligands for each receptor can be roughly divided in two groups based on the shape of the tails of the distributions. The first group (Figures 8A and 9A) is composed of relatively larger and multiply substituted ligands characterized by slower-varying tails with larger probabilities at low binding energies ($u < -15$ kcal/mol). The second group of complexes (Figures 8B and 9B) is composed of more compact ligands characterized by higher probabilities at intermediate binding energies ($-15 < u < 0$ kcal/mol) which decay rapidly with decreasing binding energy.

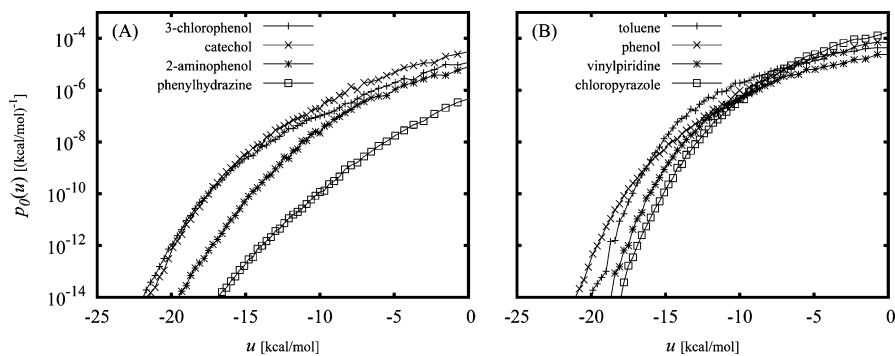


Figure 9. Favorable binding energy tails of the binding energy distributions of the L99A/M102Q T4 lysozyme complexes.

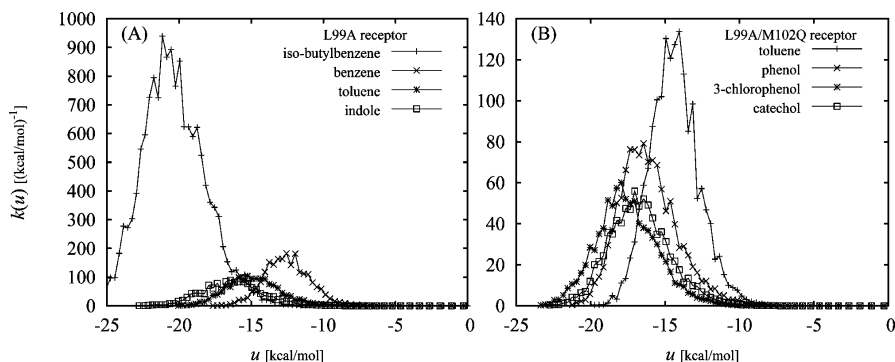


Figure 10. Binding affinity densities (eq 17) for the binders of the L99A (A) and L99A/M102Q (B) receptors.

The computed binding affinity densities $k(u)$ (eq 17) for the four binders to each receptor are shown in Figure 10. $k(u)$ measures the contribution of conformations with binding energy u to the binding constant. In these figures curves of larger magnitude correspond to the stronger binders. The range of binding energies over which $k(u)$ is significant gives an indication of the energetics of the conformations of the complex that contribute to binding. For example, it is evident from these curves that the conformations that contribute to iso-butylbenzene binding to the L99A receptor tend to have more favorable binding energies than those of benzene (approximately 7 kcal/mol less favorable on average, see Figure 10).

Figure 11 summarizes the conformational decomposition analysis for the L99A/M102Q complexes of the four binders toluene, phenol, 3-chlorophenol, and catechol. Each panel in this figure shows the macrostate binding affinity densities, $k_i(u)$ from eq 20, for the major macrostates of the ligand identified using the pitch and in-plane rotation angles θ_n and θ_p as described in section 2.7. The figure legend reports the fraction of the binding constant attributed to each macrostate from eq 23, which, as shown above, is also the value of the population of that macrostate in the physical complex at $\lambda = 1$. Also reported in this figure are the macrostate-specific binding free energies $\Delta F^\circ(i)$ of each macrostate computed from eq 21.

4. Discussion

The accuracy of the standard binding free energies of the T4 lysozyme complexes obtained from BEDAM (Table 1) are comparable to the corresponding results obtained through double-decoupling calculations with explicit

solution.^{23,53,58,59,72} The method correctly discriminates the binders from the nonbinders for the set of compounds we examined. As pointed out above, the values of the BEDAM binding free energy estimates are systematically smaller in magnitude than the experiments. The fact that for most complexes the estimates for the L99A receptor are offset by a constant amount suggests that the systematic error is due in part to overhydration of the apo receptor rather than to other effects, such as ligand–receptor interactions or ligand hydration, which are dependent on ligand size and ligand composition. The ligand-free L99A hydrophobic cavity is not occupied by water molecules.⁷³ Our implicit solvent model, however, assumes that the cavity is filled with high dielectric and does not sufficiently penalize hydration sites within hydrophobic enclosures. We suspect that the hydration free energy for the unbound L99A receptor is overly stabilizing, thereby disfavoring binding.

The data for the polar L99A/M102Q receptor suggests a more complex origin for the errors in computed affinities. The calculated binding free energy of toluene to the L99A/M102Q receptor (Table 1) is in better agreement with the experiment (0.7 kcal/mol difference) than for the L99A receptor. This indicates that the model error originating from the overhydration of the apo L99A/M102Q receptor is smaller than for the apo L99A receptor, conceivably because the former is simply more hydrated.⁵⁵ The remainder of the errors for the L99A/M102Q receptor ligands vary from ligand to ligand and are probably due to overly weak ligand–receptor interactions since AGBNP2 hydration free energies generally do not appear to systematically overestimate hydration free energies of small molecules.⁴¹ Incomplete sampling of ligand and receptor conformations can also be a source of errors.

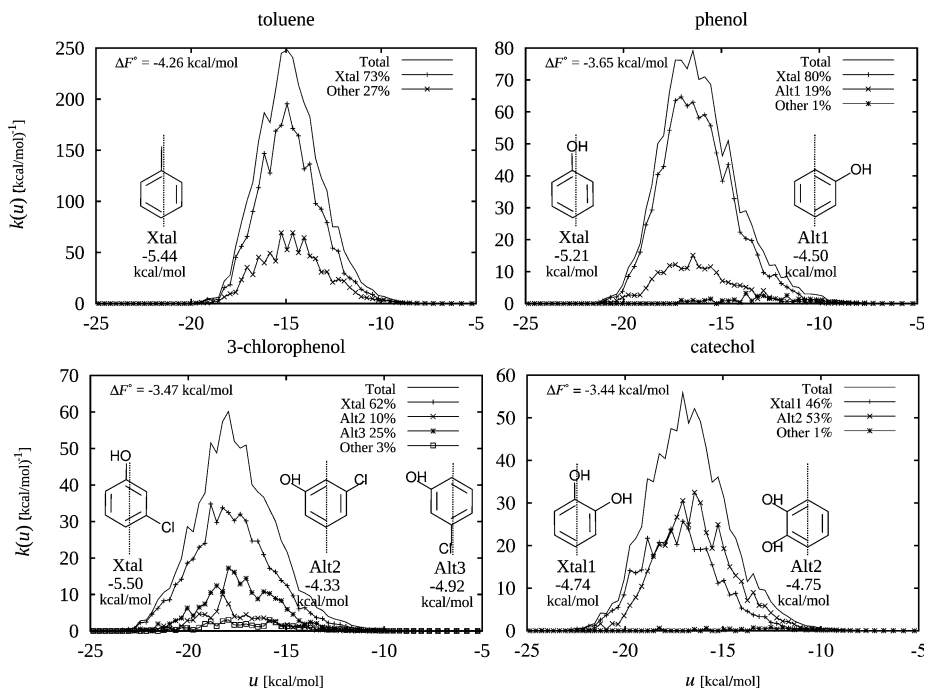


Figure 11. Conformational decomposition of the binding affinity densities for the binders of the L99A/M102Q receptor (toluene, phenol, 3-chlorophenol, and catechol). Ligand conformational macrostates labeled “Xtal” correspond to conformations observed crystallographically; other states are labeled as “Alt”. The catch-all macrostate, which includes any conformation not included in the definition of any of the other states, is labeled as “Other”. Representative conformations of the ligand for each macrostate are schematically shown in the insets; the dotted line represents the orientation within the binding site of the crystallographic conformation. The macrostate-specific binding free energy of each macrostate from eq 21 is reported below the representative conformation. The binding affinity densities $P_0(\lambda)k_\lambda(u)$ for each macrostate (eq 20), weighted by the respective populations at $\lambda = 0$, are shown such that they sum to the total binding affinity density (eq 19). The relative contribution (eq 23) of each macrostate to the overall binding constant is indicated as a percentage in the legend.

We observed, for example, particularly slow convergence of the binding free energy of 3-chlorophenol probably due to the multiple, and nearly degenerate, ligand poses for this ligand (see below and Figure 11).

The magnitude and shape of the low binding energy tail of the binding energy probability distributions $p_0(u)$ presented in Figures 8 and 9 aid in the rationalization of the trends observed in the binding free energies. In general, higher probabilities in this range of binding energies is reflected in more favorable binding free energies. For example, in Figure 8A the curves for the two binders, iso-butylbenzene and indole, lie well above those for the two nonbinders, *tert*-butylbenzene and trimethylbenzene. In addition, because of the exponential weighting in the integral for the binding constant (eq 12), low binding energies have a larger effect on the binding affinity than intermediate binding energies. This explains in part why, for example, toluene and phenol bind the L99A/M102Q receptor better than 4-vinylpyridine and 4-chloro-1h-pyrazole (Figure 9B).

The 16 ligands we investigated can be classified in two groups based on the shape of their binding energy distributions. Bulkier ligands (Figures 8A and 9A) correspond to binding energy distributions that extend to lower binding energies and that tend to have smaller probabilities at intermediate binding energies than those of more compact ligands (Figures 8B and 9B). This behavior is consistent with the interpretation that larger ligands are capable of forming stronger interactions with the receptor but only in specific

poses that occur with low probability. Conversely, small ligands can achieve favorable binding affinity by means of larger numbers of conformations with intermediate binding energies. The role of these two modes of binding can also be seen by comparing the distributions of related ligands. For example, the distributions for benzene and toluene (Figure 8B), which bind the L99A receptor with similar affinity, reveal that addition of a methyl group substituent has a small effect on the binding affinity because the gain in the strength of ligand–receptor interactions ($u < -15$ kcal/mol) is almost completely counterbalanced by the loss of probability density at intermediate binding energies ($-15 < u < 0$ kcal/mol), which is explained by the fewer alternative ways for toluene to properly fit into the binding site compared to benzene.

The distributions also help explain differences in binding affinities between related ligands with distributions of similar shape. For example, the shapes of the distributions for catechol, a binder for the L99A/M102Q receptor, and 2-aminophenol, a nonbinder, are very similar (Figure 9A), indicating a similar pattern of interactions for the two ligands. The probability tail for catechol, however, is down shifted by about 2 kcal/mol, an amount that mirrors the difference between their binding free energies (Table 1). These results suggest that the lower binding affinity of 2-aminophenol is energetic in origin. Analysis of the binding energy terms (eq 8) indeed indicates that the two ligands differ mainly in the desolvation free energy term which opposes binding of

2-aminophenol by approximately 2 kcal/mol more than catechol. Analogously, the binding energy distributions of phenol bound to the L99A and L99A/M102Q receptors are similar in shape (Figures 8B and 9B) with the one for the L99A/M102Q receptor down shifted by approximately 4 kcal/mol relative to the other. This energy shift, caused by the hydrogen-bonding interaction between phenol and Q102, is responsible for the better affinity of phenol for the L99A/M102Q receptor compared to the L99A receptor.

Some of the computed binding affinity densities [$k(u)$, from eq 17] are shown in Figure 10. These functions measure the contribution of conformations with binding energy u to the binding constants, given by the areas under the curves. We see from these figures that the range of binding energies that contribute to binding varies significantly from one ligand to another. For example, the binding affinity density for iso-butylbenzene is significant for binding energies around -20 kcal/mol (Figure 10A), whereas $k(u)$ for benzene is significant only for much less favorable binding energies (approximately 7 kcal/mol less favorable on average). In contrast, the difference of the binding free energies between these two ligands is much smaller (about 1.2 kcal/mol, Table 1). The reason for this is that although iso-butylbenzene can form strong interactions with the receptor, it can do so with low probability (from Figure 8A at $u \approx -20$ kcal/mol, $p_0(u) \approx 10^{-11}$ kcal/mol $^{-1}$). In contrast, benzene achieves favorable binding by means of more numerous conformations of moderate binding energies; for instance, at $u \approx -13$ kcal/mol the probability density for benzene is approximately $p_0(u) \approx 10^{-6}$ kcal/mol $^{-1}$ (Figure 8B), a value 5 orders of magnitude greater than for iso-butylbenzene above.

These probabilistic effects, which oppose binding, can be quantified by the residual $\Delta F_{\text{AB}}^{\circ} - \langle u \rangle_1$ between the binding free energy and the average binding energy $\langle u \rangle_1 = \langle V_1 - V_0 \rangle_1$ at $\lambda = 1$. This quantity can be expressed as the sum of the conformational entropy loss, $\Delta S_{\text{conf}}^{\circ}$,¹⁵ and the reorganization energy, ΔE_{reorg} , upon binding given by

$$-T\Delta S_{\text{conf}}^{\circ} = \Delta F_{\text{AB}}^{\circ} - \Delta E_{\text{AB}} = \Delta F_{\text{AB}}^{\circ} - (\langle V_1 \rangle_1 - \langle V_0 \rangle_0) \quad (33)$$

and

$$\Delta E_{\text{reorg}} = \langle V_0 \rangle_1 - \langle V_0 \rangle_0 \quad (34)$$

where $\Delta E_{\text{AB}} = \langle V_1 \rangle_1 - \langle V_0 \rangle_0$ is the effective enthalpy of binding and V_{λ} , given by eq 25, is the λ -dependent effective potential. Using the computed binding free energy values from Table 1 and the average binding energy values from Table 2 we obtain values for $\Delta F_{\text{AB}}^{\circ} - \langle u \rangle_1$ of 15.2 kcal/mol for iso-butylbenzene, compared to only 8.5 kcal/mol for benzene. The large difference between these residuals indicates that iso-butylbenzene, in addition to losing more conformational entropy than benzene, also induces significantly more receptor strain. Indeed, we observed that in the $\lambda = 1$ trajectory of the complex with iso-butylbenzene that the V111 residue together with helix F of the receptor are shifted away from the binding pocket compared to the complex with benzene. The positioning of these elements in the simulation of the complex with iso-butylbenzene is

Table 2. Lowest and Average Binding Energies and Corresponding Ligand Rankings

molecule	rank (calcd) ^a	min (u) ^b	min rank ^c	$\langle u \rangle_1$ ^d	$\langle u \rangle_1$ rank ^e
L99A apolar cavity					
iso-butylbenzene	1	-27.3	1	-20.4	1
benzene	2	-17.8	7	-12.5	7
toluene	3	-20.2	5	-14.8	5
indole	4	-22.9	4	-16.3	4
tert-butylbenzene	5	-24.7	2	-18.4	2
cyclohexane	6	-19.6	6	-14.1	6
1,3,5-trimethylbenzene	7	-22.9	3	-16.8	3
phenol	8	-17.5	8	-11.7	8
rank order CC ^f		0.36		0.36	
L99A/M102Q polar cavity					
toluene	1	-20.0	6	-14.8	6
phenol	2	-21.4	5	-16.1	3
3-chlorophenol	3	-23.5	2	-17.6	1
catechol	4	-22.9	3	-16.7	2
4-vinylpyridine	5	-18.7	7	-13.6	7
4-chloro-1h-pyrazole	6	-18.7	8	-12.4	8
2-aminophenol	7	-22.9	4	-15.0	5
phenylhydrazine	8	-26.2	1	-15.7	4
rank order CC ^f		-0.21		0.38	

^a Ligand rankings based on the calculated binding free energies (from Table 1). ^b Lowest binding energy found over the conformations sampled from the HREM simulation. ^c Ligand rankings based on lowest binding energy values. ^d Average binding energy at $\lambda = 1$. ^e Ligand rankings based on average binding energy values. ^f Rank order correlation coefficients between lowest/average binding energy rankings (fourth and sixth columns, respectively) and binding free energy rankings (second column).

similar to the corresponding crystal structure⁷³ except for the rotameric state of V111 which remains in the starting apo configuration instead of adopting the one seen in the crystal structure. Explicit modeling of this conformational change has been shown to improve the agreement with the experimental binding free energies.⁵⁹

We see from the computed $k(u)$ functions (Figure 10) that, as expected, the contribution from conformations with unfavorable binding energies ($u > 0$) is negligible. (This is true for both binders and nonbinders, although only the binding affinity densities of binders are shown in Figure 10.) Interestingly, this analysis shows that conformations with very favorable binding energies also contribute little to binding. For example, the smallest binding energy we observed for phenol bound to the L99A/M102Q receptor is -21.4 kcal/mol. However, as the binding affinity density for phenol shows (Figure 10B), conformations with binding energies in this low range provide a negligible contribution to the binding constant. This is because they occur with insufficient probability in the bound complex to make a difference. Consequently, it is apparent that for an accurate computation of the binding constant it is not necessary to sample binding energies well below values that are frequently found for the complex at room temperature.

Another notable and common feature of the binding affinity densities we obtained (Figure 10) is their relatively large widths, indicating that conformations with a wide range of binding energies are contributing to binding. For example, we see (Figure 10A) that the binding affinity of iso-butylbenzene is the result of appreciable contributions from conformations with binding energies in a 10 kcal/mol range

from -25 to -15 kcal/mol. In addition, conformational decomposition analysis (see discussion below and Figure 11) shows that in this system energetic heterogeneity is accompanied by extensive conformational heterogeneity.

The conformational decomposition analysis of the binding affinity densities (summarized in Figure 11 for the L99A/M102Q complexes) illustrates the wide range of ligand poses that give rise to the calculated binding free energies, even for these simple ligands with very few internal degrees of freedom. We see that in none of the cases examined is all of the binding affinity due to a single macrostate of the complex. In the case of catechol two distinct poses contribute equally to the binding affinity, and therefore, missing one of them would underestimate the binding constant by a factor of 2. Phenol presents a less extreme case in which 80% of the affinity is accounted for by the macrostate corresponding to the crystallographic pose. For toluene and 3-chlorophenol a variety of ligand poses contribute appreciably to binding in addition to the crystallographic pose. Because, as noted above, the relative contributions to binding of ligand macrostates are equal to their relative populations at $\lambda = 1$, information about these contributions can in principle be obtained from a conventional simulation of the complex. As previously noted,⁵³ however, due to kinetic trapping it is challenging in practice to achieve equilibrium between conformational macrostates without resorting to enhanced sampling strategies, like, for example, HREM.

The conformational decomposition analysis yields macrostate-specific standard binding free energies, $\Delta F^\circ(i)$ (eq 21), which correspond to the binding free energies that would be measured if ligand conformations were restricted to within specific macrostates. Macrostate-specific binding free energies have been previously introduced to compute standard binding free energies from multiple free energy calculations each focused on a single macrostate.^{53,66} The macrostate-specific binding free energies for the binders of the L99A/M102Q receptor computed in this work are reported in Figure 11. Notably, the magnitudes of macrostate-specific binding free energies often exceed that of the total binding free energy. For example, the binding free energy for the crystallographic macrostate of phenol is -5.21 kcal/mol compared to the total computed standard binding free energy of -3.65 kcal/mol. This is due to the fact that macrostate-specific binding free energies ignore the entropic loss due to the many other orientations of the ligand in solution which cannot form favorable interactions with the receptor. These effects are encoded in the populations, $P_0(i)$, of ligand macrostates in solution that, when properly combined in eq 22 with their respective macrostate-specific binding constants, yield the total binding constant.

The energetic and conformational heterogeneities we observed for the complexes studied in this work (Figures 10 and 11) illustrate why it is difficult to correlate the properties of a single conformation of the complex to the binding affinity. The binding affinity originates from multiple and diverse conformations whose contributions depend on the balance between their binding energy and their probability of occurrence. In addition, we note that empirical scoring functions for binding⁷⁴ are often applied to energetically

optimized conformations that do not necessarily contribute significantly to binding. To illustrate this point we show in Table 2 the ligand rankings for each receptor based on the most favorable binding energies observed in the simulations together with their correlations with the free energy rankings. We see that there is very little correlation in this system between the lowest binding energies and the binding free energies, particularly for the L99A/M102Q receptor for which phenylhydrazine (the poorest binder) is predicted to be the best binder based on the lowest binding energy. The average binding energies collected at $\lambda = 1$ (which correspond, for example, to the binding energy term of the single-trajectory MM-PBSA method⁷⁵) are somewhat better correlated with the binding free energies (Table 2).

Similar to docking and scoring approaches,² BEDAM is based on computing receptor–ligand interaction energies. Rather than doing so on a single or few selected ligand poses, however, in BEDAM the probability distributions of binding energies are collected from thousands of conformations drawn from canonical conformational ensembles computed with physical models of molecular interactions. The latter feature is in common with end point approaches, such as MM-PB/GBSA⁷⁶ and mining minima methods,¹⁴ which employ separate models for the binding enthalpy and binding entropy. In contrast, BEDAM is essentially a binding free energy model that, as discussed, in principle includes all enthalpic and entropic effects through the $p_0(u)$ binding energy distribution.

BEDAM bears some relationship to both potential of mean force and double-decoupling methods for computing standard binding free energies.^{8,18,21} Since they share the same statistical mechanics foundation (i.e., eq 1), in principle BEDAM yields equivalent results to these methods (to the extent that the implicit solvent models reflect the solvent potential of mean force as accurately as explicit solvation). Potential of mean force methods obtain the binding free energy by computing the free energy profile for transferring the ligand from solution to the binding site region. Similarly, the binding energy considered by BEDAM for each conformation of the complex represents the change in potential of mean force (with implicit solvation) for moving the ligand from the solution to a particular position and orientation in the binding site at fixed receptor and ligand conformations. Conformational, translational, and rotational entropic contributions are included in BEDAM by means of the exponential averaging (eq 12) of the binding energies over all possible conformations and positioning of the ligand relative to the receptor. Potential of mean force methods capture the same contributions by means of a thermodynamic cycle involving restraining and releasing steps.²⁰

Similar to double-decoupling strategies,^{8,18} BEDAM is based on an alchemical transformation to link the bound and unbound states of the complex. There are, however, conceptual and methodological differences between BEDAM and double-decoupling strategies. BEDAM is based on binding energy values computed with implicit solvation, whereas double decoupling has been employed so far only with explicit solvation. The implicit solvent representation in BEDAM makes it possible to compute binding energy

distributions that, as illustrated above, represent receptor–ligand complex fingerprints which are useful for analysis of binding interactions and their conformational decomposition. On the operational side, BEDAM involves only one simulation leg rather than two (one for the unbound ligand and one for the complex) with double decoupling. This feature is potentially advantageous for more rapid convergence of the binding free energies of highly polar and charged ligands, which, in double-decoupling and end point approaches, are the result of a nearly complete cancellation between the large free energies of the solvated and bound states.²⁰ Because binding energies are averaged over a single simulation, BEDAM results will be less sensitive to statistical errors. Care should be taken, however, to achieve the correct balance between interatomic and hydration interactions for charged groups with implicit solvation.⁴¹

Other notable operational differences between BEDAM and double-decoupling approaches, as commonly implemented,^{23,77} involve the free energy computational protocol and treatment of restraints. In double decoupling the free energy of turning off the interactions of the ligand from its environment is conducted in a series of steps, or windows, evaluated independently or sequentially by second generation FEP free energy estimators.^{65,78} In BEDAM instead the binding free energy is computed through the binding energy distribution using a strategy based on Hamiltonian replica exchange (HREM) umbrella sampling and histogram reweighting (WHAM). In potential of mean force and double-decoupling implementations the thermodynamic path connecting the bound and unbound states is commonly divided into a series of intermediate steps involving imposition and removal of conformational restraints and the separate decoupling and recoupling of electrostatic, van der Waals, and steric interactions.⁷⁷ In BEDAM no conformational restraints are imposed other than those pertinent to the definition of the bound complex as prescribed by the theory. In addition, in this work we have not found it necessary to decouple electrostatic interactions separately from other interactions. This is due in part to the fact that BEDAM interactions are not completely turned on or off. Rather, as λ goes from 0 to 1, ligand–solvent interactions are smoothly replaced by ligand–receptor interactions. Similarly, the sampling efficiency gained by Hamiltonian replica exchange partly explains the ability of BEDAM to reach reasonable convergence in the simple systems considered here without imposing tight restraints or subdividing the calculation across multiple conformational states.^{53,66}

One of the key features of BEDAM is the close match between the underlying theory and its numerical implementation. Indeed, the HREM umbrella sampling and WHAM protocols are particularly well suited for the computation of binding energy distributions on which BEDAM is based. HREM in λ space allows for the rapid equilibration between stable conformations of the complex, which provide the energetic driving force for binding, and for efficient coverage of the families of conformations not as suitable for binding, which provide the entropic cost of association. HREM MD trajectories are not limited to a single λ step; rather they can explore the whole range of the thermodynamic path, thereby

enhancing conformational sampling and mixing. At the same time, conformational sampling is focused in the binding site region, thereby avoiding spending computing time to sample uninteresting regions of conformational space that do not contribute to the binding free energy. The ladder of λ values for HREM can be chosen so that uniform coverage of the range of binding energies important for binding is achieved.

The WHAM reweighting procedure applies naturally to the computation of the binding energy distribution at $\lambda = 0$ from the binding energy values extracted from the HREM trajectories. Through WHAM, each sample contributes to the overall free energy result and not only to the λ value at which it was collected. Furthermore, the dynamic range of binding energy probabilities that can be robustly probed with this method can be very large, thereby enhancing the reliability of the binding free energies computed from it. This is because the relative precision of the computed binding energy distribution $p_0(u)$ depends mainly on the number of samples collected at binding energy u , rather than the value of $p_0(u)$ itself. The multistate Bennett acceptance ratio method (MBAR),⁴⁷ which does not require binning, could be equivalently used in BEDAM to compute the binding free energy by reweighting. However, the computation of the binding energy distributions $p_0(u)$ and $p_1(u)$, which are useful analytical tools, require binning. Another strategy that could prove convenient in cases where the extreme favorable binding energy tail of $p_0(u)$ is difficult to sample, is to adopt a parametric model for $p_0(u)$ whose parameters are optimized from the collected samples by means of inference analysis.⁹ Future work will address these potential enhancements for BEDAM.

5. Conclusions

We presented the binding energy distribution analysis method (BEDAM) for calculation of protein–ligand standard free energies of binding with implicit solvation. We have shown that the theory underlying the method is homologous to the test particle insertion method of solvation thermodynamics with the solute–solvent potential replaced by the effective binding energy of the protein–ligand complex. Accordingly, in BEDAM the binding constant is computed by means of a Boltzmann-weighted integral (eq 12) of the probability distribution of the binding energy obtained in the canonical ensemble in which the ligand, while positioned in the binding site, is embedded in the solvent continuum and does not interact with receptor atoms. We have shown that the binding energy distribution encodes all of the physical effects of binding and that its analysis yields useful insights into energetic and entropic contributions to binding. We have also shown how joint probability distributions can be constructed to perform the conformational decomposition of the computed binding affinity.

We developed an efficient computational protocol for the binding energy distribution based on the AGBNP2 implicit solvent model, parallel Hamiltonian replica exchange sampling, and histogram reweighting. We have shown that the sampling of ligand conformations is such that the results are independent of the starting conformation of the complex. We have also confirmed that the results are converged with

respect to the definition of the binding site volume. Illustrative results are reported for a set of known binders and nonbinders of the L99A and L99A/M102Q mutants of T4 lysozyme receptor. The method is found to be able to correctly discriminate the known binders from the known nonbinders. The computed standard binding free energies of the binders are found to be in reasonably good agreement with reported calorimetric measurements. The conformational decomposition analysis of the results reveals that the binding affinities of these systems reflect contributions from multiple binding modes spanning a wide range of binding energies.

Despite the positive results for the T4 lysozyme model system, further work will be needed to apply the BEDAM method to more complex targets. Systems of pharmaceutical interest often involve larger and more flexible ligands as well as more flexible receptors than those studied here. Ligand and receptor reorganization make important contributions to the binding affinity,^{8,79–82} and although these effects are implicitly included in the theory, they are only partially accounted for by the BEDAM conformational sampling protocol as currently implemented. The HREM conformational sampling algorithm developed here is quite successful in sampling a variety of receptor–ligand poses, but it is expected to be less useful for ergodic sampling at physiological temperature of the internal degrees of freedom of some ligands.⁸³ Similarly, the current conformational sampling methodology has not been designed to extensively sample receptor conformations. While acceptable for the relatively rigid receptors investigated in this study, this limitation will pose a significant challenge for application of the method to other kinds of receptors. To address these issues we are currently exploring the applicability of methods based on the combination of Hamiltonian and temperature replica exchange to enhance conformational sampling. The robustness of implicit solvent modeling is also of concern for the general applicability of the method. The two major challenges are probably the accurate modeling of charged groups and treatment of structural water molecules.⁸⁴ Further tests of the method will likely offer useful insights into the improvement of the representation of the solvent for these applications.

Acknowledgment. We are grateful to Dr. Michael K. Gilson for critically reviewing the manuscript and elucidating some aspects of the theory. We are also grateful to three anonymous reviewers for helpful suggestions and insights. This work has been supported in part by a research grant from the National Institute of Health (GM30580). The calculations reported in this work have been performed at the BioMaPS High Performance Computing Center at Rutgers University funded in part by the NIH shared instrumentation grant no. 1 S10 RR022375.

References

- Jorgensen, W. L. *Science* **2004**, *303*, 1813–1818.
- Guvench, O.; MacKerell, A. D. *Curr. Opin. Struct. Biol.* **2009**, *19*, 56–61.
- Brooijmans, N.; Kuntz, I. D. *Annu. Rev. Biophys. Biomol. Struct.* **2003**, *32*, 335–373.
- McInnes, C. *Curr. Opin. Chem. Biol.* **2007**, *11*, 494–502.
- Shoichet, B. K. *Nature* **2004**, *432*, 862–865.
- Zhou, Z.; Felts, A. K.; Friesner, R. A.; Levy, R. M. *J. Chem. Inf. Model.* **2007**, *47*, 1599–1608.
- Gilson, M. K.; Zhou, H.-X. *Annu. Rev. Biophys. Biomol. Struct.* **2007**, *36*, 21–42.
- Mobley, D. L.; Dill, K. A. *Structure* **2009**, *17*, 489–498.
- Free Energy Calculations. In *Theory and Applications in Chemistry and Biology*; Chipot, C., Pohorille, A., Eds.; Springer Series in Chemical Physics; Springer: Berlin, Heidelberg, 2007.
- Zhou, H.-X.; Gilson, M. K. *Chem. Rev.* **2009**, *109*, 4092–4107.
- Tembe, B. L.; McCammon, J. A. *Comput. Chem.* **1984**, *8*, 281.
- Shirts, M.; Mobley, D.; Chodera, J. *Ann. Rep. Comput. Chem.* **2007**, *3*, 41–59.
- Jorgensen, W. L.; Thomas, L. L. *J. Chem. Theory Comput.* **2008**, *4*, 869–876.
- Chang, C.-E.; Gilson, M. K. *J. Am. Chem. Soc.* **2004**, *126*, 13156–13164.
- Chang, C. A.; Chen, W.; Gilson, M. K. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 1534–1539.
- Kollman, P. A.; Massova, I.; Reyes, C.; Kuhn, B.; Huo, S.; Chong, L.; Lee, M.; Lee, T.; Duan, Y.; Wang, W.; Donini, O.; Cieplak, P.; Srinivasan, J.; Case, D. A.; Cheatham, T. E. *Acc. Chem. Res.* **2000**, *33*, 889–897.
- Chong, L. T.; Pitera, J. W.; Swope, W. C.; Pande, V. S. *J. Mol. Graph. Model.* **2009**, *27*, 978–982.
- Gilson, M. K.; Given, J. A.; Bush, B. L.; McCammon, J. A. *Biophys. J.* **1997**, *72*, 1047–1069.
- Lee, M. S.; Olson, M. A. *Biophys. J.* **2006**, *90*, 864–877.
- Woo, H.-J.; Roux, B. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6825–6830.
- Deng, Y.; Roux, B. *J. Phys. Chem. B* **2009**, *113*, 2234–2246.
- Hermans, J.; Subramaniam, S. *Isr. J. Chem.* **1986**, *27*, 225–227.
- Boyce, S. E.; Mobley, D. L.; Rocklin, G. J.; Graves, A. P.; Dill, K. A.; Shoichet, B. K. *J. Mol. Biol.* **2009**, *394*, 747–763.
- Chen, J.; Brooks, C.; Khandogin, J. *Curr. Opin. Struct. Biol.* **2008**, *18*, 140–148.
- Felts, A. K.; Gallicchio, E.; Chekmarev, D.; Paris, K. A.; Friesner, R. A.; Levy, R. M. *J. Chem. Theory Comput.* **2008**, *4*, 855–868.
- Zhang, Y. *Curr. Opin. Struct. Biol.* **2009**, *19*, 145–155.
- Scheraga, H. A.; Khalili, M.; Liwo, A. *Annu. Rev. Phys. Chem.* **2007**, *58*, 57–83.
- Felts, A. K.; Harano, Y.; Gallicchio, E.; Levy, R. M. *Proteins: Struct., Funct., Bioinf.* **2004**, *56*, 310–321.
- Felts, A.; Andrec, M.; Gallicchio, E.; Levy, R. In *Water and Biomolecules-Physical Chemistry of Life Phenomena*; Springer Science: New York, 2008.
- Gallicchio, E.; Zhang, L. Y.; Levy, R. M. *J. Comput. Chem.* **2002**, *23*, 517–529.
- Mobley, D.; Chodera, J.; Dill, K. *J. Phys. Chem. B* **2008**, *112*, 938–946.

- (32) Shoichet, B. K.; Leach, A. R.; Kuntz, I. D. *Proteins* **1999**, *34*, 4–16.
- (33) Majeux, N.; Scarsi, M.; Apostolakis, J.; Ehrhardt, C.; Caflisch, A. *Proteins* **1999**, *37*, 88–105.
- (34) Maple, J. R.; Cao, Y.; Damm, W.; Halgren, T. A.; Kaminski, G. A.; Zhang, L. Y.; Friesner, R. A. *J. Chem. Theory Comput.* **2005**, *1*, 694–715.
- (35) Huang, N.; Kalyanaraman, C.; Irwin, J. J.; Jacobson, M. P. *J. Chem. Inf. Model.* **2006**, *46*, 243–253.
- (36) Naim, M.; Bhat, S.; Rankin, K. N.; Dennis, S.; Chowdhury, S. F.; Siddiqi, I.; Drabik, P.; Sulea, T.; Bayly, C. I.; Jakalian, A.; Purisima, E. O. *J. Chem. Inf. Model.* **2007**, *47*, 122–133.
- (37) Carlsson, J.; And er, M.; Nervall, M.;  qvist, J. *J. Phys. Chem. B* **2006**, *110*, 12034–12041.
- (38) Su, Y.; Gallicchio, E.; Das, K.; Arnold, E.; Levy, R. *J. Chem. Theory Comput.* **2007**, *3*, 256–277.
- (39) Michel, J.; Essex, J. W. *J. Med. Chem.* **2008**, *51*, 6654–6664.
- (40) Gallicchio, E.; Levy, R. *J. Comput. Chem.* **2004**, *25*, 479–499.
- (41) Gallicchio, E.; Paris, K.; Levy, R. M. *J. Chem. Theory Comput.* **2009**, *5*, 2544–2564.
- (42) Su, Y.; Gallicchio, E. *Biophys. Chem.* **2004**, *109*, 251–260.
- (43) Levy, R. M.; Zhang, L. Y.; Gallicchio, E.; Felts, A. K. *J. Am. Chem. Soc.* **2003**, *125*, 9523–9530.
- (44) Roux, B.; Simonson, T. *Biophys. Chem.* **1999**, *78*, 1–20.
- (45) Sugita, Y.; Okamoto, Y. *Chem. Phys. Lett.* **1999**, *314*, 141–151.
- (46) Gallicchio, E.; Andrec, M.; Felts, A. K.; Levy, R. M. *J. Phys. Chem. B* **2005**, *109*, 6722–6731.
- (47) Shirts, M. R.; Chodera, J. D. *J. Chem. Phys.* **2008**, *129*, 124105.
- (48) Sugita, Y.; Kitao, A.; Okamoto, Y. *J. Chem. Phys.* **2000**, *113*, 6042–6051.
- (49) Rick, S. W. *J. Chem. Theory Comput.* **2006**, *2*, 939–946.
- (50) Hritz, J.; Oostenbrink, C. *J. Chem. Phys.* **2008**, *128*, 144121.
- (51) Woods, C. J.; Essex, J. W.; King, M. A. *J. Phys. Chem. B* **2003**, *107*, 13703–13710.
- (52) Jiang, W.; Hodoscek, M.; Roux, B. *J. Chem. Theory Comput.* **2009**, *5*, 2583–2588.
- (53) Mobley, D. L.; Chodera, J. D.; Dill, K. A. *J. Chem. Phys.* **2006**, *125*, 084902.
- (54) Eriksson, A. E.; Baase, W. A.; Wozniak, J. A.; Matthews, B. W. *Nature* **1992**, *355*, 371–373.
- (55) Wei, B. Q.; Baase, W. A.; Weaver, L. H.; Matthews, B. W.; Shoichet, B. K. *J. Mol. Biol.* **2002**, *322*, 339–355.
- (56) Morton, A.; Baase, W. A.; Matthews, B. W. *Biochemistry* **1995**, *34*, 8564–8575.
- (57) Graves, A. P.; Brenk, R.; Shoichet, B. K. *J. Med. Chem.* **2005**, *48*, 3714–3728.
- (58) Deng, Y.; Roux, B. *J. Chem. Theory Comput.* **2006**, *2*, 1255–1273.
- (59) Mobley, D. L.; Graves, A. P.; Chodera, J. D.; McReynolds, A. C.; Shoichet, B. K.; Dill, K. A. *J. Mol. Biol.* **2007**, *371*, 1118–1134.
- (60) Boresch, S.; Tettinger, F.; Leitgeb, M.; Karplus, M. *J. Phys. Chem. B* **2003**, *107*, 9535–9551.
- (61) Widom, B. *J. Phys. Chem.* **1982**, *86*, 869–872.
- (62) Beck, T. L.; Paulaitis, M. E.; Pratt, L. R. *The Potential Distribution Theorem and Models of Molecular Solutions*; Cambridge University Press: New York, 2006.
- (63) Pohorille, A.; Pratt, L. R. *J. Am. Chem. Soc.* **1990**, *112*, 5066–5074.
- (64) Widom, B. *J. Chem. Phys.* **1963**, *39*, 2808–2812.
- (65) Lu, N.; Singh, J. K.; Kofke, D. A. *J. Chem. Phys.* **2003**, *118*, 2977–2984.
- (66) Jayachandran, G.; Shirts, M. R.; Park, S.; Pande, V. S. *J. Chem. Phys.* **2006**, *125*, 084901.
- (67) Mihailescu, M.; Gilson, M. K. *Biophys. J.* **2004**, *87*, 23–36.
- (68) Lu, N.; Kofke, D. A. *J. Chem. Phys.* **2001**, *114*, 7303–7311.
- (69) Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L. *J. Phys. Chem. B* **2001**, *105*, 6474–6487.
- (70) Feig, M.; Brooks, C. *Curr. Opin. Struct. Biol.* **2004**, *14*, 217–224.
- (71) Banks, J.; et al. *J. Comput. Chem.* **2005**, *26*, 1752–1780.
- (72) Mobley, D. L.; Chodera, J. D.; Dill, K. A. *J. Chem. Theory Comput.* **2007**, *3*, 1231–1235.
- (73) Morton, A.; Matthews, B. W. *Biochemistry* **1995**, *34*, 8576–8588.
- (74) Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P. *J. Comput. Aided Mol. Des.* **1997**, *11*, 425–445.
- (75) Brown, S. P.; Muchmore, S. W. *J. Chem. Inf. Model.* **2007**, *47*, 1493–1503.
- (76) Simonson, T.; Archontis, G.; Karplus, M. *Acc. Chem. Res.* **2002**, *35*, 430–437.
- (77) Wang, J.; Deng, Y.; Roux, B. *Biophys. J.* **2006**, *91*, 2798–2814.
- (78) Shirts, M. R.; Bair, E.; Hooker, G.; Pande, V. S. *Phys. Rev. Lett.* **2003**, *91*, 140601.
- (79) Sherman, W.; Day, T.; Jacobson, M. P.; Friesner, R. A.; Farid, R. *J. Med. Chem.* **2006**, *49*, 534–553.
- (80) Yang, C.-Y.; Sun, H.; Chen, J.; Nikolovska-Coleska, Z.; Wang, S. *J. Am. Chem. Soc.* **2009**, *131*, 13709–13721.
- (81) DeLorbe, J. E.; Clements, J. H.; Teresk, M. G.; Benfield, A. P.; Flake, H. R.; Millspaugh, L. E.; Martin, S. F. *J. Am. Chem. Soc.* **2009**, *131*, 16758–16770.
- (82) Lapelosa, M.; Arnold, G. F.; Gallicchio, E.; Arnold, E.; Levy, R. M. *J. Mol. Biol.* **2010**, *397*, 752–766.
- (83) Okumura, H.; Gallicchio, E.; Levy, R. M. *J. Comput. Chem.* **2010**, *31*, 1357–1367.
- (84) Abel, R.; Young, T.; Farid, R.; Berne, B. J.; Friesner, R. A. *J. Am. Chem. Soc.* **2008**, *130*, 2817–2831.

JCTC

Journal of Chemical Theory and Computation

Dynamics-Based Discovery of Allosteric Inhibitors: Selection of New Ligands for the C-terminal Domain of Hsp90

Giulia Morra,^{⊥,†} Marco A. C. Neves,^{⊥,†} Christopher J. Plescia,[‡] Shinji Tsustsumi,[§]
Len Neckers,[§] Gennady Verkhivker,^{||} Dario C. Altieri,[‡] and Giorgio Colombo^{*,†}

Istituto di Chimica del Riconoscimento Molecolare, CNR, Via Mario Bianco 9, 20131 Milano, Italy, Department of Cancer Biology, University of Massachusetts Medical School, Worcester, Massachusetts 01605, Urologic Oncology Branch, Center for Cancer Research, National Cancer Institute, Bethesda, Maryland 20892, and Department of Pharmaceutical Chemistry, School of Pharmacy and Center for Bioinformatics, University of Kansas, Lawrence, Kansas 66047

Received June 17, 2010

Abstract: The study of allosteric functional modulation in dynamic proteins is attracting increasing attention. In particular, the discovery of new allosteric sites may generate novel opportunities and strategies for drug development, overcoming the limits of classical active-site oriented drug design. In this paper, we report on the results of a novel, *ab initio*, fully computational approach for the discovery of allosteric inhibitors based on the physical characterization of signal propagation mechanisms in proteins and apply it to the important molecular chaperone Hsp90. We first characterize the allosteric “hot spots” involved in interdomain communication pathways from the nucleotide-binding site in the N-domain to the distal C-domain. On this basis, we develop dynamic pharmacophore models to screen drug libraries in the search for small molecules with the functional and conformational properties necessary to bind these “hot spot” allosteric sites. Experimental tests show that the selected molecules bind the Hsp90 C-domain, exhibit antiproliferative activity in different tumor cell lines, while not affecting proliferation of normal human cells, destabilize Hsp90 client proteins, and disrupt association with several cochaperones known to bind the N- and M-domains of Hsp90. These results prove that the hits alter Hsp90 function by affecting its conformational dynamics and recognition properties through an allosteric mechanism. These findings provide us with new insights on the discovery and development of new allosteric inhibitors that are active on important cellular pathways through computational biology. Though based on the specific case of Hsp90, our approach is general and can readily be extended to other target proteins and pathways.

Introduction

The dynamic properties of proteins play key roles in all aspects of protein functions, ranging from molecular recognition and binding to enzymatic activity.¹ A better knowledge

of dynamics from experiments and theory makes it now feasible to model the conformational properties of several proteins at the atomic scale.² Functional dynamics is determined by a complex interplay of covalent and noncovalent interactions that define the relative population of three-dimensional (3D) structures (determined by their free energies) and the possible interconversion kinetic pathways among them (determined by the heights of the free energy barriers between them).^{3,4} Binding of a ligand or substrate at an active site or of a protein partner at a certain region of the structure may select specific accessible conformations endowed with specific functional properties.⁵

* Corresponding author. E-mail: giorgio.colombo@icrm.cnr.it.
Telephone: ++39-02-28500031.

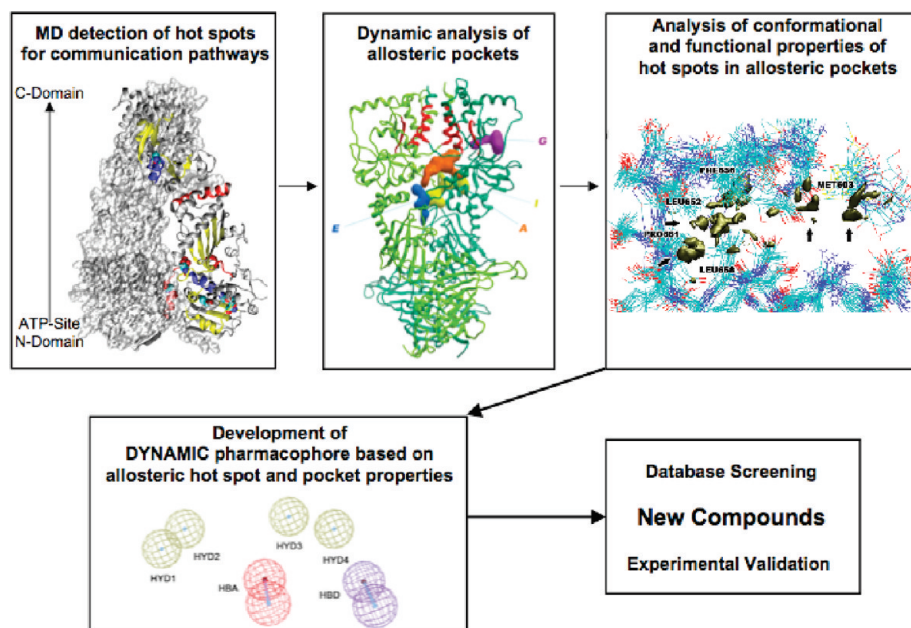
[⊥] These two authors contributed equally to this work.

[†] Istituto di Chimica del Riconoscimento Molecolare.

[‡] University of Massachusetts Medical School.

[§] National Cancer Institute.

^{||} University of Kansas.

Scheme 1. Computational Biology Strategy for the Selection of Allosteric Inhibitors Exploiting Protein Dynamics^a

^a From the analysis of the dynamics of the activated state of the protein (Hsp90 bound to ATP), the hot spot residues active in mediating signal transduction are identified. Analysis of possible binding pockets centered on these residues identifies putative allosteric binding sites. A consensus model of functional interactions with the hot spots together with structural shape constraints is used for pharmacophore modeling. The ensemble-based approach ensures the incorporation of receptor flexibility into the pharmacophore model. Small molecule databases are then screened for leads fitting with the pharmacophoric hypothesis, and selected hits are tested experimentally.

Allosteric molecular perturbations may alter the covalent and noncovalent forces that determine the fine combination of dynamic modes at the basis of molecular recognition and function. This reverberates in a modification of the protein's structural and/or dynamic properties, causing a response where a specific function can be switched on, fine-tuned, regulated, or blocked. Perturbation of the protein's conformational ensemble may be achieved through several mechanisms, including ligand binding, covalent modifications, or mutations. A variation of the protein state at a certain site may thus impact on the binding affinity in a distal functional region, such as the active site or a protein contact surface.⁶ Information transmission between distant functional sites in proteins represents a manifestation of nonlocal interactions between residues.

The molecular mechanisms of these site-to-site communication phenomena are of great interest, especially since understanding the dynamic connectivity that favors signaling through structures may reveal new allosteric binding sites and illuminate molecular mechanisms of functional regulation.⁷ Moreover, achieving these goals would offer tremendous opportunities in the design of new drugs, protein engineering, and chemical genomics. Rational targeting of alternative sites may reveal new chemotypes for potential inhibitors and offer new strategies to interfere with protein–protein interactions, which are generally recognized as challenging targets.⁸

In this paper, we present a novel rational strategy for the computational-based discovery of allosteric inhibitors of molecular chaperones. In particular, we aim to perturb the functions of the activated form of the chaperone heat shock protein-90 (Hsp90), by the rational selection of antagonists

with the structural and functional characteristics necessary to target “hot-spot” allosteric residues located in the C-terminal domain (CTD), which are dynamically coupled to the N-terminal ATP binding-site and may potentially affect Hsp90 chaperone function. To this end, we build on the results of a long-range coordination analysis that we developed to study Hsp90 communication pathways and dynamics at atomic resolution.⁹ Our approach showed that conformational changes and coordination between the N- and C-domains are responsive to specific nucleotide binding. This propagates molecular signals long-range, selectively to functionally important residues and secondary structure elements at the CTD, which define possible allosteric binding sites. The physicochemical properties of newly detected functional CTD sites are used here to build receptor-based 3D pharmacophore models. This allows us to identify novel antagonists of Hsp90 chaperone function that target a site distant from the active site, inhibit several important protein–protein interactions, and show the ability to interrupt biological pathways important for cancer cell proliferation (Scheme 1). As expected the activities of the molecules, selected from a publicly available database, do not make them immediate candidates for drug development. However, it is important to underline that our main goal is to use information on the protein dynamics to identify potential hits. Drug design and medicinal chemistry efforts can then be started on this basis to improve the activities and pharmacokinetic properties of our hits.

We focus on Hsp90 as an example of a molecular system in which ligand-based activation and signal communication between physically distant domains underlies protein–protein interactions and biological function. Hsp90 is a homodimer

in which each protomer is characterized by a modular architecture with three domains: an N-terminal regulatory domain (NTD), a middle domain (M-domain), and a carboxy-terminal domain (CTD).^{10–12} The biological activity of Hsp90 depends on ATP binding and hydrolysis, which is coupled to a conformational cycle that involves the opening and closing of a dimeric molecular clamp formed by the association of the NTDs of Hsp90.¹³ In solution, the protein exists as a dimer owing to the stable association of highly conserved motifs in its CTD. ATP binds to the NTDs of Hsp90, stabilizes their transient dimerization,¹³ and sends a conformational signal to the CTD, which is responsible for the acquisition of the ATP-ase competent conformation required for chaperone activity. Moreover, the Hsp90 chaperone function is finely regulated in the cell by physical association with a number of cochaperones that regulate the ATP-ase activity or direct Hsp90 to interact with different client proteins. Different cochaperones bind to different domains of Hsp90 (for a complete review see).¹⁴

Hsp90 has a well-established role in the conformational maturation, stability, and function of a wide range of “client” proteins within the cell. In cancer cells, Hsp90 is overexpressed and intersects signaling pathways essential for tumor maintenance, and its inhibition through drugs targeting the N-terminal ATP-site showed promising therapeutic perspectives.¹⁵

The results presented here open the possibility to rationally expand the chemical space of Hsp90 antagonists to effective inhibitors of allosteric communications.

Experimental Section

In this section, we describe the computational and experimental procedures in detail. In the Computational Details Section, we first describe how molecular dynamics (MD) simulations and signal communication modeling were carried out. These experimental details have already been fully described in.⁹ Here we are reporting them for clarity. Next we describe the development of the pharmacophore modeling, virtual ligand screening (VLS), the docking of known and new compounds to the newly discovered pockets.

In the Experimental Procedures Section, we report on the experimental procedures used to test the small molecules.

Computational Details. MD Simulations. The MD simulation trajectories used in this work were carried out as already described in ref 9. The details and the full description of the MD set up and runs can thus be found in the published paper dealing with the characterization of the ligand modulation of Hsp90 dynamics.⁹

Briefly, the crystal structure (pdb entry 2CG9)¹⁰ containing yeast Hsp90 dimer bound to ATP was employed as a starting point for the simulations. The system was solvated in a tetrahedral solvation box contains around 57 000 particles. All simulations and the analysis of the trajectories were performed using the GROMACS software package¹⁶ using the GROMOS96 force field¹⁷ and the SPC water model.¹⁸

The ATP-bound Hsp90 dimer system was first energy relaxed with 2000 steps of steepest-descent energy minimization followed by another 2000 steps of conjugate gradient

energy minimization. The energy minimization was used to remove possible bad contacts from the initial structures. The system was then equilibrated by a 50 ps of MD run with position restraints on the protein and ligand to allow relaxation of the solvent molecules. The first equilibration run was followed by a second 50 ps run without position restraints on the solute. The first 5 ns of the trajectory was not used in the subsequent analysis in order to minimize convergence artifacts. Equilibration of the trajectory was checked by monitoring the equilibration of quantities, such as the root-mean-square deviation (rmsd) with respect to the initial structure, the internal protein energy, and fluctuations were calculated on different time intervals. The electrostatic term was described by using the particle mesh Ewald algorithm. The LINCS¹⁹ algorithm was used to constrain all bond lengths. For the water molecules, the SETTLE algorithm²⁰ was used. A dielectric permittivity, $\epsilon = 1$, and a time step of 2 fs were used. All atoms were given an initial velocity obtained from a Maxwellian distribution at the desired initial temperature of 300 K. The density of the system was adjusted performing the first equilibration runs at *NPT* condition by weak coupling to a bath of constant pressure ($P_0 = 1$ bar, coupling time $\tau_P = 0.5$ ps).²¹ In all simulations, the temperature was maintained close to the intended values by weak coupling to an external temperature bath²¹ with a coupling constant of 0.1 ps. The proteins and the rest of the system were coupled separately to the temperature bath. The structural cluster analysis was carried out using the method described by Daura and co-workers with a cutoff of 0.25 nm.²²

Signal Propagation Analysis. This approach was also already described in ref 9. It is based on the adaptation of a recent approach proposed by Bahar and co-workers to the analysis of all-atom MD simulation trajectories. The analysis of signal propagation, which was developed based on elastic network models,²³ defines signal transduction events in proteins as directly related to the fluctuation dynamics of atoms, defining the communication propensities (CP) of a pair of residues as a function of the fluctuations of inter-residue distances. Residues whose C α –C α distance fluctuates with a relatively small intensity during the trajectory are supposed to communicate more efficiently than residues whose distance fluctuations are large. In the former case, a perturbation at the one site, affecting the C α position, is likely to be visible (reflected) at the second site, while in the latter case, the communication is less efficient due to the intrinsic amplitude of the distance fluctuations. The CP of any two residues is defined as the mean-square fluctuation of the interresidue distance defining $d_{ij} = |\vec{r}_i - \vec{r}_j|$ as distance between the C α atoms of residues i and j , respectively:

$$CP = \langle (d_{ij} - d_{ij,ave})^2 \rangle$$

By projecting these quantities on the 3D structures of the protein bound to different ligands, it will be possible to identify possible differences in the interdomain and inter-protomer long-range redistributions of interactions.

The CP was calculated for any pair of residues during the trajectory. It is worth noting that CP describes the distance

fluctuation of the two residues, therefore, low CP values characterize residues that move in a highly coordinated fashion and hence may be involved in the efficient relay of conformational signals.⁹ The average CP value for consecutive amino acids along the sequence, calculated considering for each residue i the neighbors comprised between $i - 4$ and $i + 4$, is 0.025. The average CP value for residues distant more than 40 Å is 0.12. In the presence of ATP, around 1% of residue pairs have CP < 0.025 even if they are at distances larger than 40 Å.⁹ Therefore, in the presence of ligands, a number of very distant residues may have a low CP value and display high coordination despite their physical separation, and we set CP = 0.025 as a convenient threshold for discriminating high dynamic coordination at long distance. CP values at increasing distances were scanned through histogram analysis. Each bin of the histogram refers to a residue and gives the fraction of residues that have high coordination with it (CP < 0.025) at distances larger than an increasing cutoff of 40, 60, and 80 Å, respectively. Residues corresponding to histogram peaks define regions that are specifically involved in efficient long-range correlations. Results of the analysis are fully reported in ref 9.

Upon increase of the residue–residue distance in the CP scanning histograms, some peaks become progressively smaller or disappear, since the fraction of effectively coordinated residues decreases at longer physical distances. On the other hand, since the total number of possible pairs also decreases with increasing distance, for some residues the fraction of highly coordinated partners may grow at longer distances, and those residues we define to be strongly active in long-range signaling.

The residues in the C-terminal preserving the most efficient communication propensities with the ATP site were used to define the possible allosteric-binding pocket. They comprise the NTD residues 81–95 and 121–140 (Hsp90 residues numbering as in the pdb entry 2CG9) that have a long-range signaling propensity with segments 574–580 and with the two C-terminal interface helices, made of residues 645–654 (helix 4) and 661–671 (helix 5), respectively.

MD-Based Pharmacophore Modeling. Hsp90 dimer conformations were collected at every 0.5 ns of the final 20 ns MD trajectory using the GROMACS software package and superimposed at the putative C-terminal binding pocket, i.e., residues 475–477, 591–595, 602–603, and 652–657 of one monomer and residues 502–504, 591–595, and 656–662 of another. Superimposition was performed based on backbone atoms. GREATER v1.2.2, the graphical user interface for GRID v22a, was used to calculate molecular interaction fields (MIFs) with the probes DRY (hydrophobic), O (sp² carbonyl oxygen) and N1 (neutral flat amide NH).²⁴ The protein was considered rigid and a 31 × 27 × 18 Å grid box was centered at the binding pocket. Grid spacing was set to 0.25 Å. Local energy minima, defined as isocontours from probes DRY (−0.8), O (−7), and N1 (−7 kcal/mol), were represented with the VMD software v1.8.6.²⁵ Binding pocket regions with consistently favorable interactions along the MD trajectory were used to define 3D pharmacophore features of a pharmacophore hypothesis for Hsp90 C-terminal binding.

Table 1. Pharmacophore Conformational Properties^a

	HYD1	HYD2	HYD3	HYD4	HBA
HYD2	2.6				
HYD3	9.9	7.9			
HYD4	13.4	11.7	4.5		
HBA	9.8	8.1	5.3	5.6	
HBD	15.8	14.1	7.0	2.6	7.1

^a Distance constraints in Å.

Pocket analysis was also carried out with the PocketFinder module of the ICM suite.²⁶ Probe atoms (carbon, oxygen, and nitrogen atoms) were placed at the center of higher density areas and converted into a pharmacophore hypothesis using the Catalyst ViewHypothesis workbench of the Catalyst v4.1.1 software. Local energy minima identified with the DRY, O, and N1 probes were converted into hydrophobic and hydrogen-bond acceptor and donor features, respectively. Flexibility was taken into account with 1.6 Å radius tolerances around each pharmacophore feature, i.e., spherical volumes where matching chemical groups should be located. Projection points from which the extended hydrogen-bond partner participates, i.e., Arg591 and/or Ser657 hydrogen bonding an acceptor group and Asp503 and/or Ser602 hydrogen bonding a donor group, were created in order to mimic the location of their side chains during the MD simulation. Shape filtering was done by filling the common binding cavity along the last 10 ns MD trajectory with chemical probes (carbon atoms) and converting them into inclusion volumes using the convert molecule to shape tool of the Catalyst ViewHypothesis workbench. The minimum similarity tolerance was set to 0.5.

The final pharmacophore hypothesis consisted of a 3D arrangement of six features (i.e., four hydrophobic regions and one each hydrogen-bond acceptor and donor) located at defined positions. These were surrounded by 1.6 Å radius tolerance spheres, assessing the area in space that should be matched by corresponding chemical functions of the virtual screening molecules. The hydrogen-bond acceptor and donor features additionally include a vector indicating the direction of the interaction. The desirable shape of the new virtual screening hits was delimited by a series of inclusion volumes. Table 1 reports on the distance constraints for the pharmacophore model generated.

Virtual Screening. The NCI database was downloaded from the 2007 release of the ZINC library²⁷ and converted into a multiconformer Catalyst database. A maximum of 100 conformations, within a 20 kcal/mol energy range above the calculated global minimum, were generated for each molecule using the “FAST” conformational analysis model of catDB utility program. The pharmacophore hypothesis was screened using the “fast flexible database search” settings.

Docking of Known and Newly Discovered Compounds. Initial models for novobiocin, its derivatives, and the selected small molecules described in the paper were generated using the standard building blocks of MAESTRO v8.5 and minimized with MACROMODEL v8.1,^{28,29} using the Merck molecular force field (MMFF),³⁰ the Polak–Ribiere conjugate gradient (PRCG) minimization method with an energy convergence criterion of 0.05 kJ/mol and the generalized

Born equation/surface area (GB/SA)³¹ continuum solvation model with parameters for water (dielectric constant $\epsilon = 78$). Five thousand steps of the systematic unbounded multiple minimum (SUMM) method implemented in MACRO-MODEL were used in order to allow a full exploration of the conformational space. Autodock Tools v1.5 was used to prepare ligands and receptors for docking, namely, to remove water molecules, add hydrogens, compute Gasteiger charges,³² and merge nonpolar hydrogens. Side chain charges were assigned according to their pK_a . Blind docking experiments on the whole Hsp90 C-terminus domain were performed with the novobiocin derivatives using AutoDock v4.0.³³ Grid maps were generated with AutoGrid v4.0 using a 0.375 Å grid spacing. The Lamarckian genetic algorithm was employed for all docking runs. An initial population of 150 individuals randomly placed on the Hsp90 C-terminus domain was created. Random orientations and torsions were used. The number of generations was set to 25 million, and the maximum number of energy calculations was set to 27 thousands. A mutation rate of 0.02 and a crossover rate of 0.8 were used, and the local search frequency was set up at 0.06. Two hundred independent runs were performed for each compound with the parameters described above. Results differing by less than 2 Å in positional rmds were clustered together and represented by the result with the most favorable free energy of binding.

Initial geometries for the virtual screening hit compounds were collected from the ZINC database.²⁷ Docking runs were limited to the allosteric binding pocket at the dimer interface.

Experimental Procedures. *Cell Viability, Elisa Tests, and Akt Folding. Cells and Cell Cultures.* Human prostate adenocarcinoma PC3 and lung adenocarcinoma H460 cells were obtained from the American Type Culture Collection (ATCC, Manassas, VA) and maintained in cultures as recommended by the supplier. Human umbilical vein endothelial cells (HUVEC) were obtained from Clonetics. Rat A10 smooth muscle cells were the generous gift of Dr. Michael Conte, University of California, San Francisco.

Antibodies. Antibodies to b-actin (Sigma-Aldrich) and Akt (CST, Inc., Danvers, MA) were used.

Binding Assays. Plastic microtiter wells were coated with increasing concentrations (0–150 μ M) of the various compounds, blocked in 3% gelatin, and further mixed with recombinant full length Hsp90 or Hsp90 C-domain (residues 629–732, 1 mg/mL) produced in BL-21 *E. Coli* as a GST fusion protein, and further isolated from the GST frame by thrombin cleavage. After a 2 h incubation at 22 °C, compound binding under the various conditions tested was detected with an antibody to Hsp90, followed by a peroxidase-conjugated secondary reagent and quantification of absorbance at 405 nm.

Cell Viability Analysis. The various normal or tumor cell types (2×10^5 /ml, 50 mL) were seeded in triplicates in 96-well plates and incubated with increasing concentrations of the various Hsp90-C terminus compounds (0–150 mM) for 16 h at 22 °C. At the end of the incubation, cultures were analyzed for cell viability by an 3(4,5-dimethyl-thiazoyl-2-yl)2,5 diphenyl-tetrazolium bromide (MTT) colorimetric assay with absorbance at 405 nm. In other experiments,

tumor cell types were incubated with various concentrations of Hsp90 C-terminus compounds, and whole cell extracts were analyzed by Western blotting.

Statistical Analysis. Data were analyzed using the two-sided unpaired *t* test on a GraphPad software package for Windows (Prism). A *p* value of 0.05 was considered as statistically significant.

Cochaperone and Client Protein Interactions with Coimmunoprecipitation Assays. Cell Culture, Transfection, and Immunoprecipitation. COS7 cells (American Type Culture Collection) were cultured in a temperature-controlled incubator (37 °C and 5% CO₂) in Dulbecco's modified Eagle's medium (DMEM) medium supplemented with 10% (v/v) fetal bovine serum (FBS), 10 mM 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (HEPES, pH 7.0), 2 mM glutamine, 1 mM of sodium pyruvate, and nonessential amino acids (Biosource/Invitrogen). Cells were transiently transfected with pcDNA3 empty vector or pcDNA3 containing flag-tagged wild-type human Hsp90alpha by using FuGene6 (Roche Applied Science), following the manufacturer's instructions. Twenty-four hours after transfection, cells were treated with 100 μ M of indicated compounds for 1 h. Then, cells were lysed (20 mM HEPES, 100 mM NaCl, 1 mM MgCl₂, 0.1% NP-40, 20 mM Na₂MoO₄, phosphatase inhibitor (Roche), and protease inhibitors(Roche)), and incubated with anti-flag antibody-conjugated beads (Sigma) for 2 h at 4 °C. Coimmunoprecipitated proteins were identified by immunoblotting with indicated antibodies recognizing Flag (Affinity Bioreagents), ERBB2 (Santa Cruz), CDK4 (Santa Cruz), p60^{Hop} (Cell Signaling), p50^{Cdc37} (Neomarkers), p23 (Affinity Bioreagents), or AHA1 (Rockland). See also ref 34.

Results

Background: Hot Spots in Signal Transduction from the ATP-Site to the CTD. Different dynamic states of Hsp90 can be switched on/off in response to the presence of a specific ligand at the ATP-binding site. In this context, we generated a computational model aimed to identify the substructures (subdomains, secondary structure elements, single residues) that play a relevant role in the dynamic communication between a certain binding site and distal regions of the protein implied in function. The results showed, at atomic resolution, that the identity of mediators of the cross-talk between N- and C-domains was dependent on the specific nucleotide activating differential functional motions. Briefly, in our approach, which builds on the work of Chennubotla and Bahar,²³ the CP between any two residues, as a function of fluctuation of their distance components, is evaluated. CP describes a communication time; therefore, low CP values are related to efficiently communicating residues. The threshold for high communication efficiency is the CP value calculated for four consecutive residues along the sequence. Hot spots for signal transduction are identified by calculating for each residue the fraction of all other protein residues that have high communication efficiency with it (CP lower than the threshold) at distances larger than an increasing cutoff of 40, 60, and 80 Å. Distant,

physically separated residues that have a more efficient communication than that defined by the “local” threshold define the regions specifically involved in efficient long-range signal transduction.⁹

This analysis illuminated different pathways of signal transduction that selectively depend on the ligand identity. In particular, specific clusters of residues participate in the signal transduction from the N-terminal nucleotide-binding site to the CTD. In the ATP-bound, active form of the chaperone, long-range communication from the binding site is mainly directed to residues at the CTD interface. In particular, NTD residues 81–95 and 121–140 involved in ATP recognition (residue numbering from 2CG9.pdb) show a consistently high long-range coordination with segments 574–580 and with the C-terminal interface helices, made of residues 645–654 (helix 4) and 661–671 (helix 5), respectively (see Supporting Information, Figure S1).

Identification of Allosteric Pockets. The C-terminal interface region with higher communication propensity with the distal ATP-binding site was then subjected to structural investigation to detect potential binding sites centered on the communication hot spots. Cluster analysis of the trajectories was used to identify the most representative conformations of the CTD. Individual frames were grouped into 21 clusters, with the most populated five accounting for 84% of the structural diversity.

These representative structures and the original crystal structure (2CG9.pdb) were subjected to analysis with the pocketFinder module of the ICM software,²⁶ complemented by visual inspection. Nine potential binding pockets with volume and area suitable for interaction with drug-like compounds were identified in the X-ray crystal structure (Scheme 1, Table 2, and Figure S1b,c of the Supporting Information). Interestingly, only pocket A is consistently detected in all representative MD conformations, increasing in volume and area and defining a binding tunnel at the dimer interface suitable to accommodate small compounds able to interact directly with the hot spot residues involved in efficient long-range coordination (Figure 1a, Table 2, and Figure S1c of the Supporting Information).

Allosteric Inhibitor Discovery: Pharmacophore Modeling Based on Signal Transduction Information. Next, we used the information on signal transduction, conformational states spanned by hot spot residues, and conformational properties of pocket A together with the analysis of their chemical properties to develop pharmacophore models for virtual screening of small molecule databases. The pharmacophores are designed to recapitulate the complementary interactions necessary to guarantee productive binding with the putative allosteric site.

Structures from the final 20 ns of the MD simulations were used. Local molecular interaction fields (MIF) minima were calculated at the allosteric site with the GRID force field and probes accounting for hydrophobic (DRY) and hydrogen-bond acceptor (O) and donor (N1) interactions.³⁵ Isosurfaces at -0.8 kcal/mol derived from the DRY probe highlight four hydrophobic regions related to favorable interactions with apolar residues, such as Met 603, Leu 652, Phe 656, Leu 658, and Pro 661 (Figure 1b). Surfaces defined at an energy

Table 2. Drug-Like Binding Sites Identified with the ICM PocketFinder Module^a

structure cluster time (ns)	pockets																			
	A		B		C		D		E		F		G		H		I			
	volume	area	volume	area	volume	area	volume	area	volume	area	volume	area	volume	area	volume	area	volume	area		
X-ray $t = 0$	453.9	497.8	325.2	312.7	324.8	330.2	249.5	242.2	230.5	214.2	227.7	279.3	213.9	256.3	212.4	226.5	184.0	231.9		
1 $t = 27.9$	243.2	225.6					409.3	469.2	409.3	469.2	151.5	209.0	455.7	501.3	455.7					
2 $t = 65.2$	670.4	543.8					236.8	304.9	236.8	304.9			250.4	262.4			484.8	465.8		
3 $t = 46.3$	551.6	549.3					192.5	207.5	158.9	204.0							266.9	290.4		
4 $t = 9.5$	331.2	298.7					453.8	441.2					192.3	233.9						
5 $t = 51.1$	532.7	558.2											375.7	369.3			157.1	200.4		
residues	chain a: 477, 478, 590, 591, 593–595, 597, 599, 606, 607, 657, 658		chain b: 470, 472, 473, 507, 511, 514, 520–525, 527, 528, 582–586, 589		chain a: 470, 472, 473, 491, 514, 521–527, 581–586, 589		chain b: 472, 490, 491, 494, 496, 527–530, 532, 537, 540, 544, 568–570, 579		chain a: 598–601		chain a: 486, 573–576, 620, 621, 624, 646, 649, 650, 653		chain a: 488, 490, 491, 494, 496, 527–530, 532, 537, 540, 568–570, 579		chain a: 662, 665, 666, 669, 670		chain a: 415, 418, 422, 426, 444–446, 454, 500–503, 506		chain b: 486, 487, 573–576, 620, 621, 646, 649, 650, 653	

^a Volume (\AA^3) and area (\AA^2) are provided for each of these pockets. Blank spaces indicate either that no binding pocket was found or that the volume and area were relatively small for binding of drug-like compounds (volume <150 and/or area <200). Residues surrounding each of the binding pockets at the pdb entry 2CG9 are shown in the last row.

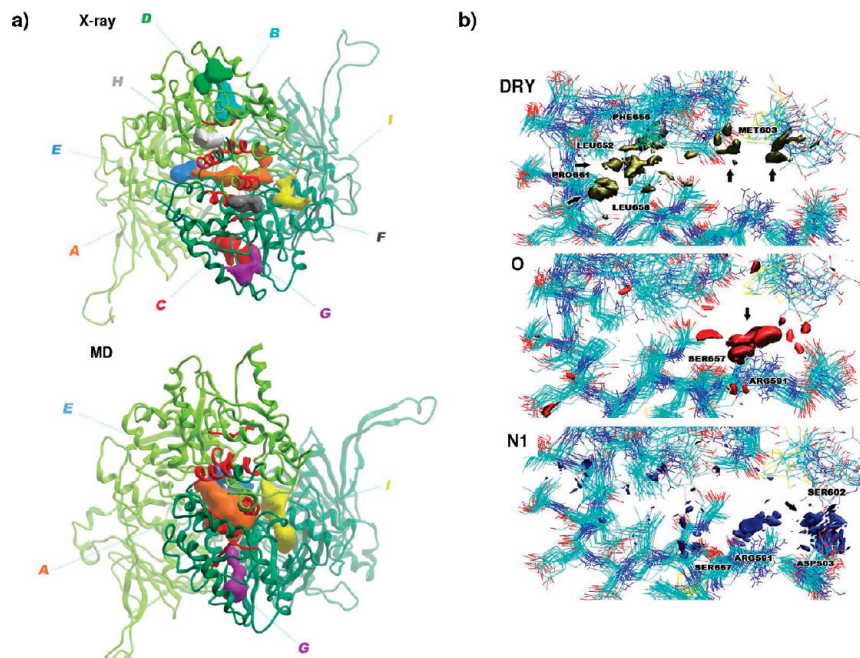


Figure 1. Pharmacophore model and resulting small molecules. (a) 3D representations of potential ligand binding pockets identified with the ICM pocketFinder module on the CTD of the Hsp90 dimer from X-ray crystal structure and from the representative structure corresponding to 65.2 ns of the MD simulation (cluster 2). Pocket A located at the dimer interface increases in area, volume, and number of contacts with the communication hot spot residues represented with a red ribbon. (b) Isosurfaces for the DRY, O, and N1 probes from the GRID force field in the putative allosteric pocket. DRY probe highlights four hydrophobic regions related to favorable interactions with apolar residues, such as Met 603, Leu 652, Phe 656, Leu 658, and Pro 661. O and N1 probes identify two regions prone to hydrogen bonding, one of these acting mainly as acceptor from Arg 591 and Ser 657 and the other as donor to Asp 503 and Ser 602, respectively.

level of -7 kcal/mol with the probes O and N1 identified two regions prone to hydrogen bonding, one of these acting mainly as acceptor from Arg 591 and Ser 657 (Figure 1b) and the other as donor to Asp 503 and Ser 602 (Figure 1b). The fluctuations in the positions, distances, and dihedral angles among the side chain functionalities of these critical residues were used to define the average and upper and lower boundaries in the positioning of the hydrogen-bond donor functions of the pharmacophore.

Taken together, these interactions defined a six-feature pharmacophore model for the virtual screening of new C-terminus targeted inhibitors of Hsp90 (Figure 2). The size and shape features of the new compounds were filtered with a set of inclusion volumes defined based on the radius and shape of pocket A at 65.2 ns of the MD simulation.

Allosteric Inhibitors: New Hits through Pharmacophore Guided Virtual Screening. The new allosteric pharmacophore model was used to perform a screening search of the NCI repository. The database contains a library of more than 290 000 compounds. Filtering of the database with the pharmacophore returned 36 hits (Figure 2), corresponding to 0.01% of the database.

Experimental Tests on Newly Discovered Hsp90 Inhibitors Targeting the C-terminal. Fourteen of the selected compounds resulting from the virtual screening could be obtained from the NCI and tested for affinity for the Hsp90 full-length protein, the CTD, for their effects on cancer and normal cell viability, for the induction of degradation of specific Hsp90 client proteins, and for their activity in disrupting Hsp90 association with cochaperones.

Molecular Interactions between Selected Molecules and Hsp90. By ELISA tests several of the discovered lead compounds (namely 6, 8, 9, 11, 12) bound the recombinant isolated Hsp90 C-domain in a specific and saturable manner (Figure 3a). Functionally, treatment of lung adenocarcinoma H460 or prostate adenocarcinoma PC3 cells with the selected compounds resulted in a concentration-dependent loss of cell viability (Figure 3b). This response was specific for inhibition of cancer-related signaling, as the implicated compounds did not reduce the viability of normal A10 smooth muscle cells or human umbilical vein endothelial cells (Figure 3b).

Selected Hits Inhibit Hsp90 Chaperone Function and Impact on Hsp90 Association with Cochaperones. We next asked whether the cytotoxic effect exerted by compounds 6, 8, 9, 11, and 12 was due to loss of Hsp90 client proteins resulting from inhibition of chaperone function. Consistent with this model, a preliminary analysis of compounds 6, 8, and 9 induced a concentration-dependent loss of the Hsp90 client protein, the kinase Akt, in tumor cells. Selected compounds were active in a concentration range between 25 and 100 μ M, with activities comparable to those of known C-terminal inhibitors. As control, compound 5, which showed no effect on tumor cell viability, did not reduce Akt levels in tumor cells (Supporting Information, Figure S2).

Selected compounds were also tested in a different experimental setting (see Materials and Methods in Supporting Information) using coimmunoprecipitation assays to probe the interaction of Hsp90 with client and cochaperone proteins. In this test, compound 6 clearly disrupted interactions with two kinase client proteins ERBB2 and CDK4.

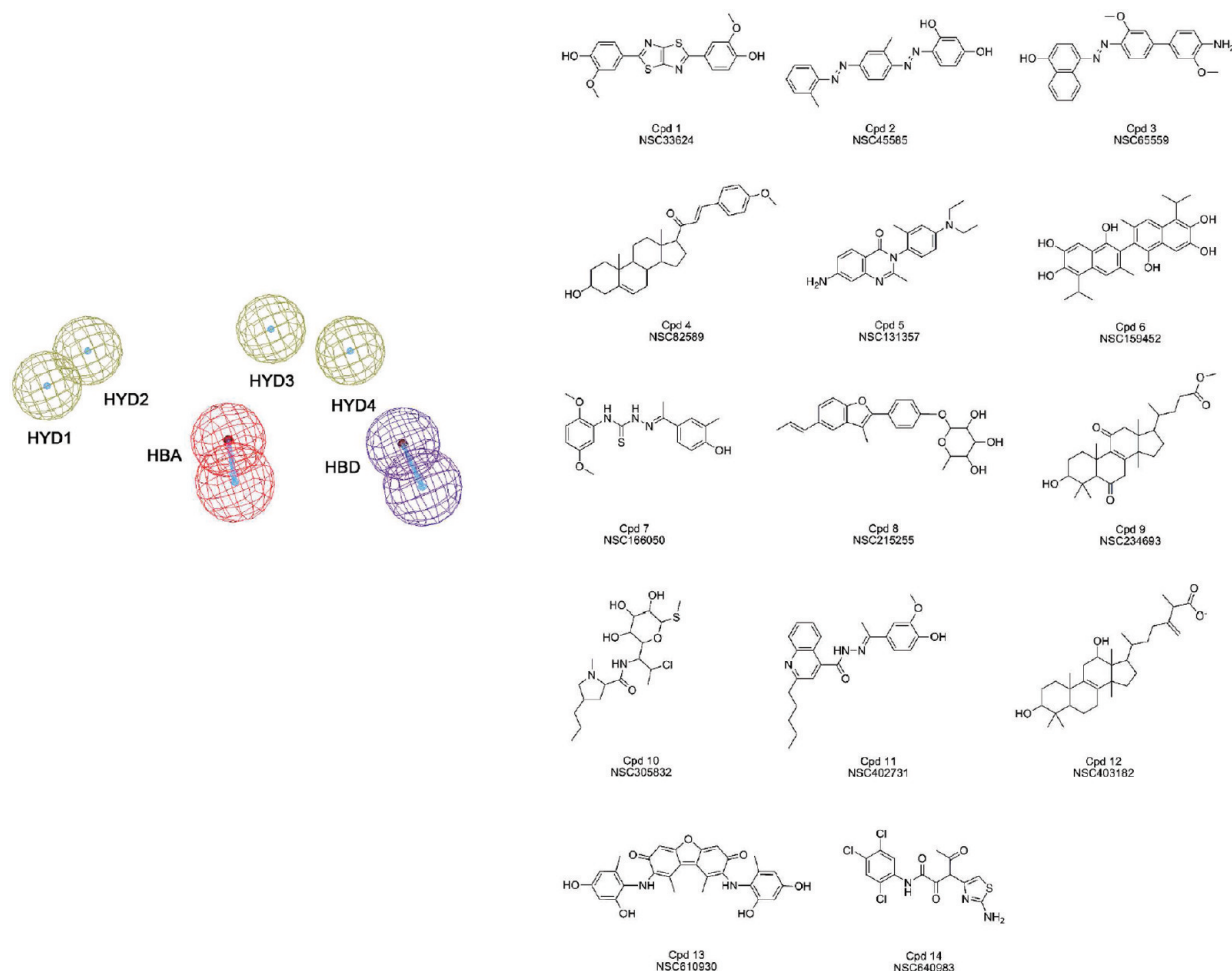


Figure 2. The pharmacophore and selected hits. The resulting six-feature pharmacophore and the molecular structures of the compounds selected from virtual screening with the pharmacophore model of the NCI database.

Moreover, compound 6 was shown to impair the binding of cochaperones p23, p50, and Aha1. While Aha1 interacts with the middle segment of Hsp90, p23 and p50 bind Hsp90 at the NTD. Since compound 6 is selected to interact with the CTD, these data suggest that it likely alters Hsp90 conformational equilibrium affecting client and cochaperone binding via an allosteric mechanism. Interestingly, compound 6 emerged as the only one able to slightly reduce Hsp90 association with p60, a cochaperone known to bind to the CTD. Binding of the inhibitor at the CTD may directly interfere with the physical binding of p60 at the same region of the protein (Figure 4).

The coimmunoprecipitation experiments also showed that compound 8 could dramatically disrupt ERBB2/Hsp90 association at 75 and 100 μM doses, supporting the validity of the computational design approach (Figure 4).

Overall, these results confirm the validity of the computational approach taking the full dynamics of the protein into account to discover new allosteric sites.

Binding Poses of the Hits in Hsp90 CTD. The molecular interactions of the compounds identified through the allosteric dynamic pharmacophore with the Hsp90 CTD were characterized via computational docking and analysis. The allosteric binding pocket is a small tunnel located at the dimer interface, delimited by residues 474–487, 502–503, 591–599, 602–603, and 652–657 from one monomer and residues

502–504, 591–595, and 656–662 from another (pocket A, Figure 1, and Figure S1, Supporting Information). Although already present in the Hsp90 crystal structure, the shape of the newly found putative site increases its binding complementarity to C-terminus inhibitors during the MD simulation, in the absence of the inhibitors. Multiple structures from the MD simulation of the ATP-complex were used as targets. This is equivalent to describing relevant representatives of the ensemble of conformational states, taking flexibility of the whole protein into account.

Binding poses and theoretical affinities were calculated and the results are reported in Table S1 of the Supporting Information. The whole CTD surface was scanned, and the active compounds were observed to dock selectively and favorably in the proposed allosteric pocket, consistent with MD simulations and pharmacophore analysis (Figure 5). These molecules show a good shape complementarity to pocket A at the dimer interface, establishing hydrogen bonds and hydrophobic contacts with the proposed allosteric hot spot residues.

Finally, we docked Hsp90 inhibitors targeting the CTD derived from the literature to the newly discovered allosteric pocket to further validate our approach through structure–activity relationships. No experimental structural information is available on complexes between CTD and these inhibitors. Novobiocin (IC_{50} 700 μM) and the more potent related

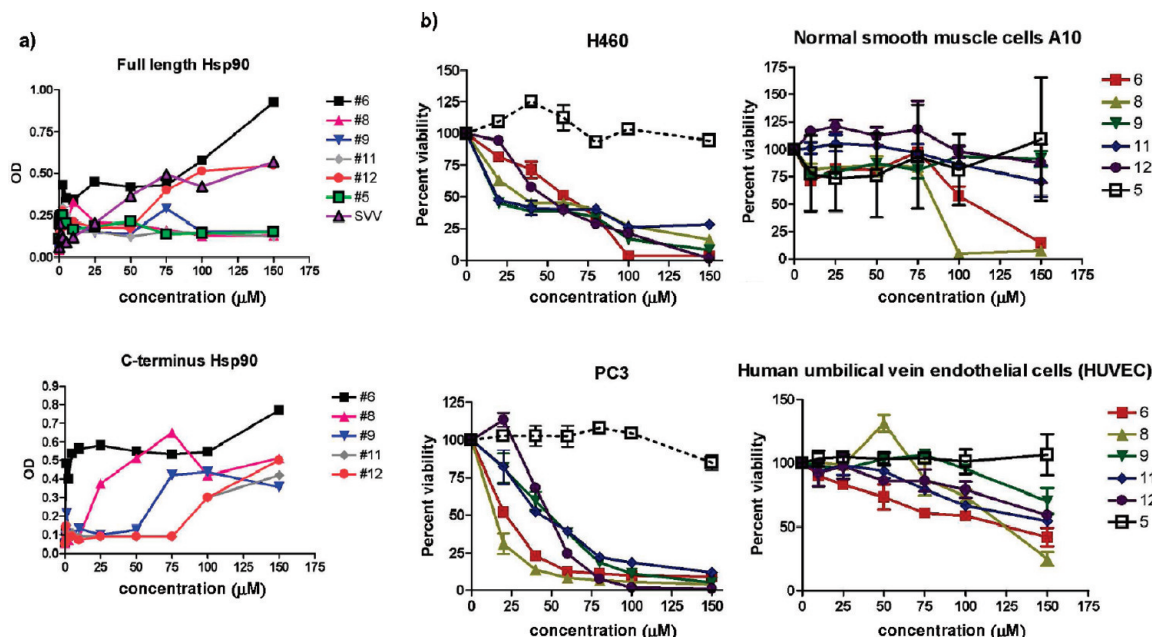


Figure 3. Small molecules bind to Hsp90 CTD and affect cancer cell viability. (a) ELISA. Microtiter wells were coated with the indicated increasing concentrations of small molecules and incubated with recombinant full-length or C-domain of Hsp90. Binding was determined using domain-specific antibodies to Hsp90 and quantified by absorbance at OD₄₀₅. Data are the mean \pm SEM of three independent experiments. (b) Inhibition of cell viability in H460 and PC3 cancer cell lines, and normal smooth muscle cells A10 and HUVEC cells, as evaluated by cell counting after a 24 h exposure to the selected small molecules. Values represent the mean (\pm SD) of three independent experiments.

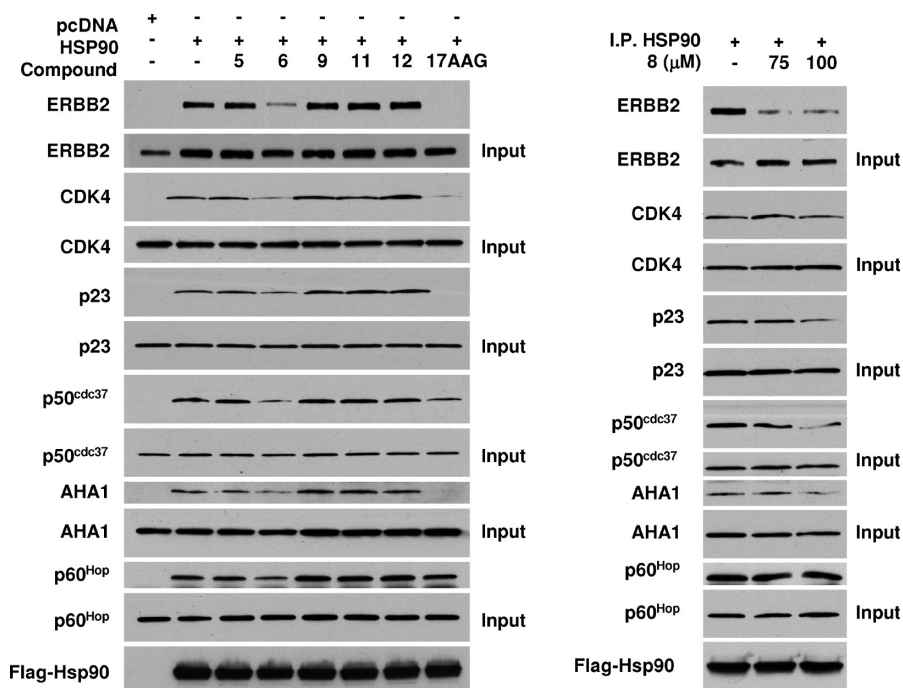


Figure 4. Inhibition of Hsp90 chaperone function. Client and cochaperone binding to Hsp90 is inhibited by small molecules 6 and 8. COS7 cells were transfected with wild-type Flag-Hsp90. After incubation, cells were treated with 100 μ M of the indicated Hsp90 inhibitor for 1 h. Then, cells were lysed, and proteins were immunoprecipitated (IP) by a Flag antibody-conjugated agarose. Indicated coprecipitating proteins were detected by immunoblotting.

derivatives ND-1 (active at 100 μ M) and ND-2 (active at 40 μ M)^{36,37} (Supporting Information, Figure S3) were thus docked to the full CTD. The calculated affinities are reported in Table S2 of the Supporting Information, along with the contacts established by the docked drugs with signal transduction hot spots. Interestingly, estimated binding energies

with novobiocin and related derivatives resulted in good agreement with their relative inhibition potencies. The strongest protein-small molecule interactions with novobiocin (-6.02 kcal/mol) and compounds ND-1 and ND-2 (-6.62 and -8.14 kcal/mol, respectively) were found with the representative structure of cluster 2 (65.2 ns frame of

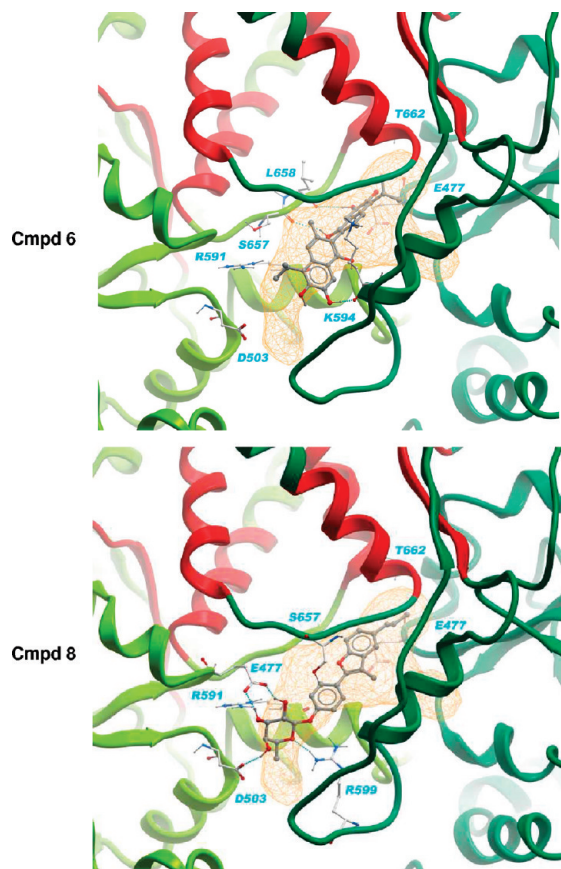


Figure 5. Binding of new hits (compounds 6 and 8) into the Hsp90 dimer. Lowest binding energies were obtained with a MD representative conformation (cluster 2, time = 65 200 ps). The protein complex is shown in a ribbon representation colored by a chain with the putative communication hot spot residues colored in red. The Hsp90 C-terminal binding pocket (pocket A) is shown as an orange line mesh and protein–ligand hydrogen bonds are represented with spheres.

the MD simulation). The three compounds make contact with residues belonging to the communication hot spot structures. Of critical importance is the disruption of a salt bridge between Glu 477 and Arg 591 after approximately 60 ns that increases the size and the volume of the binding site and improves the calculated binding affinity for compounds ND-1 and ND-2.

The qualitative good correlation between the calculated affinities and the experimental activities of the small molecules constitutes an encouraging validation of the target and of the use of information from signal transduction analysis in the detection of putative allosteric binding sites.

Discussion

The discovery of new allosteric sites may offer novel opportunities in the identification of new drugs and in the understanding of fundamental biological processes. While consensus is increasing on the importance of allosteric motions in the context of protein functional control and regulation, the relevance of using these concepts in drug design has not been fully exploited.^{8,38,39} Discovering and targeting allosteric sites can in fact lead to the expansion of the chemical space of leads and to new classes of drugs.

Most importantly, the discovery of new molecules targeted to allosteric sites may represent a viable strategy in the search for new protein–protein interactions inhibitors.⁴⁰

Protein conformational plasticity and dynamics appear to be critical for allosteric events. In the current view of allostery, a protein populates a certain ensemble of dynamic conformational states at equilibrium, and perturbations induce a shift in the relative populations of states. Signals coded by covalent or noncovalent modifications can be transmitted long-range through pre-existing pathways^{4,6} that depend on the inherent topological architecture of the protein. At a more refined level, the selection of specific communication pathways between physically separate sites depends on the fine chemical properties of the modification or the binding partner. Results from several research groups have revealed the existence of alternative interaction networks with a link to dynamic motions,^{41–44} showing that preferred relatively small, local fluctuations in proteins lead to functionally active states.

In this paper, we have built on our previous results on the atomic level characterization of the correlations between dynamics, long-range coordination, and allosteric communication between the physically distant N- and C-domains in full-length Hsp90 to develop a new strategy for computational discovery of allosteric inhibitors. Hsp90 dynamic and functional properties appeared to be highly responsive to the presence of a specific nucleotide in the ATP-site at the N-domain. Once the principal signal transduction pathways and the correlated hot spot residues that act as communication mediators between the ATP-site and the C-terminal interface have been revealed,⁹ the translation of this structural dynamical information into 3D receptor-based pharmacophore models allowed us to rationally discover new C-terminal ligands able to interfere with the chaperone function.

In this framework, we have focused on the activated form of Hsp90, which represents an important target for cancer drug discovery. Our signal transduction model revealed that the most efficient long-range communication (over >60 Å) from the binding site is mainly directed to a specific subset of residues at the CTD interface. Conformational analysis of the whole simulation trajectory showed that this site could populate a set of conformations apt to optimally accommodate known CTD targeted inhibitors. Docking of novobiocin and related derivatives to different representative protein conformations actually provided semiquantitative correlations between the activities of the compounds and their calculated binding energies.

We set out to search for new molecules targeting the newly discovered putative C-terminal allosteric site. The aim was to cause a disruptive interference in the network of interactions coding for the collective motions related to the chaperone functional activity. Our rationale was that the new hits should perturb the dynamics of the CTD substructures important for signal transduction with the NTD and interfere with the chaperone molecular recognition properties, thus disrupting association with cochaperones necessary for function and ultimately blocking client folding. To this end, we developed a pharmacophore model with complementary

functionalities for the C-terminal allosteric site, using multiple protein target structures to take the flexibility of the whole protein into account. The dynamic pharmacophore was then used to screen the NCI small molecule database. Strikingly, experimental tests proved that selected molecules bind the CTD of Hsp90. Moreover, they had important effects on the viability of two independent cancer cell lines (H460, lung e PC3, prostate), while affecting to a significantly lower degree the two normal cell types (endothelial cells and vascular smooth muscle). Compounds 6, 8, and 9 were demonstrated to inhibit Hsp90 chaperone function, as shown by the effects on the levels of Akt, an established Hsp90 client protein that requires a fully functional chaperone activity for folding and stability. Using a different experimental approach, compound 6 and 8 were confirmed to disrupt association with two more client proteins. Most importantly compound 6 was shown to affect binding to a specific subset of cochaperones, reducing the association of Hsp90 with p23, p50, Aha1, and p60.

Interestingly, p23 and p50 bind to Hsp90 NTD, while the activity of Aha1 and Akt depends on interactions with the M-domain.⁴⁵ Since the selected molecules interact directly with the C-domain, they likely alter Hsp90 molecular recognition properties by influencing its dynamics through an allosteric mechanism.

Consequently, these hits represent new leads for the development of allosteric drugs that act by tweaking the functional dynamics of the protein toward an inactive state.

The fact that molecule 6 and 8 resulted the only active hits in this second series of experiments does not exclude that the other derivatives may show similar effects under different experimental conditions. The coimmunoprecipitation assays are in fact based on Hsp90 overexpression with a drug incubation time of 1 h. Different compounds may have different binding kinetics and affinities for the chaperone, determined by specific on/off rates or different diffusion properties within the cell. The incubation time allowed for these first control experiments thus may not be sufficient to break client/cochaperone interactions.

It is worth noting at this point that compound 6 reduces the interaction between Hsp90 and the cochaperone p60, which is known to bind the CTD. Consequently, this lead also appears to perturb the molecular recognition properties of the CTD.

Importantly, the selected hits induce the disruption of Hsp90 complexes with important kinase client proteins and with cochaperones that are fundamental for Hsp90 functional activity through both the allosteric mechanisms and the abrogation of direct interactions with the CTD. This indicates that our hits act simultaneously on different biological pathways important for cancer development. It is important to underline here that the activities of our hits are still far from the ones required for efficient pharmacological applications. However, the scope of our endeavor was to identify active hits using information on an allosteric pocket obtained directly from the study of the dynamics of a complex molecular machine. Optimization of the structures through medicinal chemistry design and synthesis are currently underway.

From the applicative point of view, the possibility to rationally discover molecules that are active via allosteric and/or direct effects may facilitate the design of experiments aimed to disrupt specific interactions and to report on the behavior of the system/pathway in which the interaction is involved. All of these aspects may be important in the development of new cancer chemotherapeutics and in increasing our understanding of fundamental biochemical processes.

Moreover, we think that strategies similar to the one presented here, in which the dynamics of the target is explicitly taken into account, may be applied to the discovery of inhibitors of protein–protein interactions or of possible drug-binding sites for targets that are not easily druggable. In the former case, by carrying out an atomic resolution analysis of the protein's internal dynamics and coordination, it may be possible to isolate the interaction surfaces that are endowed with specific flexibility properties and that need specific remodeling for the molecular recognition and binding of a second protein partner. In the latter case, the knowledge of internal coordination may be exploited to identify sites where binding of a small molecule can induce the perturbation of important functional motions, resulting in the inhibition of the function of the protein or enzyme under exam.

Our findings point to several features that make approaches, such as the one presented here, attractive for the discovery and development of allosteric inhibitors of protein functions and interactions. The concept of using a combination of structural, dynamic, and long-range correlation information led us to rationally discover a new and diverse set of chemical structures with drug-like properties able to target allosteric sites very distant from the active site. In this context, we could expand the molecular diversity space of Hsp90 antagonists, selecting molecules with promising anticancer activities.

Incorporating information on functional dynamics, internal residue–residue coordination, and protein flexibility can help unveil possible binding states of the receptor that are available on the protein's energy landscape but may not be immediately evident in a single-structure representation. The discovery of alternative states can thus unveil possible allosteric binding sites, allow structurally different ligands to occupy the same site, or guide design efforts aimed at the functional and structural modification of existing leads to target-specific receptor geometries.

Overall, the use of biophysical and computational models taking dynamic and communication into account combined with pharmacophore development and screening may be useful to find new chemotypes for specific functions, to increase the yields of drug screening, and to help design new allosteric leads with important therapeutic opportunities.

Acknowledgment. This work was supported by a grant from Associazione Italiana Ricerca sul Cancro (AIRC) to G.C. G.M. gratefully acknowledges support from a “L'Oréal-Unesco for Women in Science” grant.

Supporting Information Available: Contains all the description of the materials and methods used for the calculations and experimental procedures and description of

additional tables and figures. This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- (1) Henzler-Wildman, K.; Kern, D. *Nature* **2007**, *450*, 964–972.
- (2) Smock, R. G.; Gierasch, L. M. *Science* **2009**, *324*, 198–203.
- (3) Schrank, T. P.; Bolen, D. W.; Hilser, V. J. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 16984–16989.
- (4) Hilser, V. J. *Science* **2010**, *327*, 653–654.
- (5) Boehr, D. D.; Nussinov, R.; Wright, P. E. *Nat. Chem. Biol.* **2009**, *5*, 789–796.
- (6) del Sol, A.; Tsai, C.-J.; Ma, B.; Nussinov, R. *Structure* **2009**, *17*, 1042–1050.
- (7) Hardy, J. A.; Wells, J. A. *Curr. Opin. Struct. Biol.* **2004**, *14*, 706–715.
- (8) Wells, J. A.; McClendon, C. L. *Nature* **2007**, *450*, 1001–1009.
- (9) Morra, G.; Verkhivker, G. M.; Colombo, G. *PLoS Comput. Biol.* **2009**, *5*, e1000323.
- (10) Ali, M. M. U.; Roe, S. M.; Vaughan, C. K.; Meyer, P.; Panaretou, B.; Piper, P. W.; Prodromou, C.; Pearl, L. H. *Nature* **2006**, *440*, 1013–1017.
- (11) Shiau, A. K.; Harris, S. F.; Southworth, D. R.; Agard, D. A. *Cell* **2006**, *127*, 329–340.
- (12) Dollins, D. E.; Warren, J. J.; Immormino, R. M.; Gewirth, D. T. *Mol. Cell* **2007**, *28*, 41–56.
- (13) Pearl, L. H.; Prodromou, C.; Workman, P. *Biochem. J.* **2008**, *410*, 439–453.
- (14) Zuehlke, A.; Johnson, J. L. *Biopolymers* **2010**, *93*, 211–217.
- (15) Biamonte, M. A.; Van de Water, R.; Arndt, J. W.; Scannevin, R. H.; Perret, D.; Lee, W. C. *J. Med. Chem.* **2010**, *53*, 3–17.
- (16) Van der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C. *J. Comput. Chem.* **2005**, *26*, 1701–1718.
- (17) Scott, W. R. P.; Hunenberger, P. H.; Tironi, I. G.; Mark, A. E.; Billeter, S. R.; Fennen, J.; Torda, A. E.; Huber, T.; Kruger, P.; Gunsteren, W. F. V. *J. Phys. Chem. A* **1999**, *103*, 3596–3607.
- (18) Berendsen, H. J. C.; Grigera, J. R.; Straatsma, P. R. *J. Phys. Chem.* **1987**, *91*, 6269–6271.
- (19) Hess, B.; Bekker, H.; Fraaije, J. G. E. M.; Berendsen, H. J. C. *J. Comput. Chem.* **1997**, *18*, 1463–1472.
- (20) Miyamoto, S.; Kollman, P. A. *J. Comput. Chem.* **1992**, *13*, 952–962.
- (21) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Di Nola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684–3690.
- (22) Daura, X.; Jaun, B.; Seebach, D.; Gunsteren, W. F. v.; Mark, A. E. *J. Mol. Biol.* **1998**, *280*, 925–932.
- (23) Chennubhotla, C.; Bahar, I. *PLoS Comput. Biol.* **2007**, *3*, 1716–1726.
- (24) Goodford, P. J. *J. Med. Chem.* **1985**, *28*, 849–857.
- (25) Humphrey, W.; Dalke, A.; Schulten, K. *J. Mol. Graphics* **1996**, *14*, 33–38.
- (26) Abagyan, R. A.; Totrov, M. M.; Kuznetsov, D. A. *J. Comput. Chem.* **1994**, *15*, 488–506.
- (27) Irwin, J. J.; Shoichet, B. K. *J. Chem. Inf. Model.* **2005**, *45*, 177–182.
- (28) *MacroModel*, version 8.1; Schrodinger: New York, NY, 2008.
- (29) Mohamadi, F.; Richards, N. G. J.; Guida, W. C.; Liskamp, R.; Lipton, M.; Caufield, C.; Chang, G.; Hendrickson, T.; Still, W. C. *J. Comput. Chem.* **1990**, *11*, 440–467.
- (30) Halgren, T. A. *J. Comput. Chem.* **1996**, 490–519.
- (31) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. *J. Am. Chem. Soc.* **1990**, *112*, 6127–6129.
- (32) Gasteiger, J.; Marsili, M. *Tetrahedron* **1980**, 3219–3228.
- (33) Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J. *J. Comput. Chem.* **1998**, *19*, 1639–1662.
- (34) Tsutsumi, S.; Mollapour, M.; Graf, C.; Lee, C. T.; Scroggins, B. T.; Xu, W. P.; Haslerova, L.; Hessling, M.; Konstantinova, A. A.; Trepel, J. B.; Panaretou, B.; Buchner, J.; Mayer, M. P.; Prodromou, C.; Neckers, L. *Nat. Struct. Mol. Biol.* **2009**, *16*, 1141–1147.
- (35) Neves, M. A. C.; Dinis, T. C. P.; Colombo, G.; Melo, M. L. S. *J. Med. Chem.* **2009**, *52*, 143–150.
- (36) Le Bras, G.; Radanyi, C.; Peyrat, J. F.; Brion, J. D.; Alami, M.; Marsaud, V.; Stella, B.; Renoir, J. M. *J. Med. Chem.* **2007**, *50*, 6189–6200.
- (37) Yu, X. M.; Shen, G.; Neckers, L.; Blake, H.; Holzbeierlein, J.; Cronk, B.; Blagg, B. S. J. *J. Am. Chem. Soc.* **2005**, *127*, 12778–12779.
- (38) Hardy, J. A.; Wells, J. A. *Curr. Opin. Struct. Biol.* **2004**, *14*, 706–715.
- (39) Swain, J. F.; Gierasch, L. M. *Curr. Opin. Struct. Biol.* **2006**, *16*, 102–108.
- (40) Zorn, J. A.; Wells, J. A. *Nat. Chem. Biol.* **2010**, *6*, 179–188.
- (41) Suel, G. M.; Lockless, S. W.; Wall, M. A.; Ranganathan, R. *Nat. Struct. Biol.* **2003**, *10*, 59–69.
- (42) Lee, J.; Natarajan, M.; Nashine, V. C.; Socolich, M.; Vo, T.; Russ, W. P.; Benkovic, S. J.; Ranganathan, R. *Science* **2008**, *322*, 438–441.
- (43) Swain, J. F.; Dinler, G.; Sivendran, R.; Montgomery, D. L.; Stotz, M.; Gierasch, L. M. *Mol. Cell* **2007**, *26*, 27–39.
- (44) Peng, J. W. *Structure* **2009**, *17*, 310–320.
- (45) Zuehlke, A.; Johnson, J. L. *Biopolymers* **2010**, *93*, 211–217.

CT100334N

Coarse-Grained Representations of Large Biomolecular Complexes from Low-Resolution Structural Data

Zhiyong Zhang and Gregory A. Voth*

Department of Chemistry, James Franck and Computation Institutes, University of Chicago, 5735 S. Ellis Avenue, Chicago, Illinois 60637

Received July 3, 2010

Abstract: High-resolution atomistic structures of many large biomolecular complexes have not yet been solved by experiments, such as X-ray crystallography or NMR. Often however low-resolution information is obtained by alternative techniques, such as cryo-electron microscopy or small-angle X-ray scattering. Coarse-grained (CG) models are an appropriate choice to computationally study these complexes given the limited resolution experimental data. One of the important questions therefore is how to define CG representations from these low-resolution density maps. This work provides a space-based essential dynamics coarse-graining (ED-CG) method to define a CG representation from a density map without detailed knowledge of its underlying atomistic structure and primary sequence information. This method is demonstrated on G-actin (both the atomic structure and its density map). It is then applied to the density maps of the *Escherichia coli* 70S ribosome and the microtubule. The results indicate that the method can define highly CG models that still preserve functionally important dynamics of large biomolecular complexes.

Introduction

Large biomolecular complexes are involved in many important biological processes. For example, the ribosome is a very large RNA–protein assembly that plays a central role in protein biosynthesis,^{1–3} while microfilaments^{4,5} and microtubules^{6–11} serve as structural components of the cytoskeleton.¹² It is an important but challenging task for computational biologists to simulate the large-scale functional dynamics of these biomolecular systems over the necessarily long time scales. Atomistic molecular dynamics (MD) simulation remains an important tool for studying functional dynamics of nanometer scale biomolecules (at present) up to microsecond time scales.^{13,14} However, the functional dynamics of large biomolecular complexes often occur on much longer time scales than those accessible to atomistic MD. Therefore it is necessary to perform coarse-grained (CG) modeling of these biological systems in order to simulate their long-time behaviors.^{15–19}

Generally speaking, a CG model is a reduction of the large number of degrees of freedom in an atomic structure into a

significantly smaller set. In order to establish a reasonable mapping between the atomistic and CG models, one needs to define the desired number of CG sites and then determine where to place them. We have addressed this issue previously by developing a systematic and quantitative methodology, which is particularly useful to define an aggressive CG model with a resolution lower than one site per residue for a large biomolecule.^{20,21} The method is called essential dynamics coarse graining (ED-CG), in which a CG representation is determined variationally to preserve the functional essential motion of the biomolecule.²² In the previous ED-CG implementations, the essential dynamics were characterized by principal component analysis (PCA) of an atomistic MD trajectory²⁰ or an elastic network model (ENM) of a single atomic structure.²¹ In both cases, an underlying detailed (high resolution) atomic structure of the biomolecule was needed. Furthermore, the CG sites were assumed to be contiguous along the primary sequence of the biomolecule, therefore the sequence information was also necessary.

However, it is very difficult sometimes to solve the atomic-resolution structure of a large biomolecular complex by current high-resolution experimental techniques, such as

* Corresponding author. E-mail: gavoth@uchicago.edu. Telephone: 773-702-7250.

X-ray crystallography and NMR, due to complications arising from the sheer size of such a complex and other factors, such as difficulty in crystallization (membrane complexes), etc. Instead cryo-electron microscopy (cryo-EM)^{23,24} and small-angle X-ray scattering (SAXS)^{25–27} are two experimental techniques that can obtain low-resolution models (molecular shapes) of these biomolecular complexes. There are neither atomistic details nor sequence information in these low-resolution structures, which means that the previous ED-CG methodology cannot be applied to these systems directly.

In this work, we introduce a new ED-CG scheme, which is used to define ED-CG models from a cryo-EM or SAXS structure. A technique called vector quantization (VQ) has been widely used to discretize a density map into pseudoatoms and preserve the shape of the low-resolution structure.^{28–31} It has been found that an ENM built on the pseudoatom model can describe low-frequency functional dynamics of the biomolecule quite well.^{32–35} Therefore ED-CG models are defined in the present work to capture the essential dynamics of the pseudoatom model. In the previous sequence-based ED-CG method, a CG site is a representation of a group of atoms that move together in a highly correlated fashion and are contiguous along the primary amino acid sequence at the same time. As the sequence is not directly related to the pseudoatom model, a new way of defining sites, which does not need sequence information, is developed. The new algorithm defines a space-based ED-CG model, in which a CG site, as before, represents a group of atoms that move together but are close in space instead of contiguous along the sequence. This new space-based ED-CG method can be used to define CG models of biomolecular complexes directly from their cryo-EM or SAXS structures without atomic details and sequence information. It should be noted that this method can of course be applied to atomistically detailed structures as well.

In the subsequent sections, a parameter-free ENM and the ENM-ED-CG method²¹ will first be reviewed. The new development of the space-based ED-CG method is then introduced. As a test, the resulting method is applied to the G-actin system using both its atomistic structure and a 10 Å density map, respectively. The space-based ED-CG models from the atomic structure of G-actin are also compared to the sequence-based ED-CG models. The aforementioned VQ method is a technique to define shape-based CG models from atomistic structures^{36–39} or density maps,³⁴ so the space-based ED-CG models are compared to the VQ-CG models. Two other applications of the space-based ED-CG method are to the cryo-EM density maps of the *Escherichia coli* 70S ribosome (11.5 Å) and the microtubule (8 Å), respectively, which will be compared with the VQ-CG method as well. Finally, concluding remarks are provided.

Theory and Methods

Elastic Network Model. In a typical residue-based ENM of a biomolecule,^{40–42} the positions of C_α atoms for amino acids and P atoms for nucleotides are used.^{43–46} However there is no inherent restriction on the number of atoms or residues they can represent. These CG “atoms” are connected

by effective harmonic bonds, therefore, the harmonic potential of the ENM can be written as

$$V = \sum_{i,j>i} \frac{1}{2} k_{ij} \Delta r_{ij}^2 \quad (1)$$

Here, $\Delta r_{ij} = r_{ij} - r_{ij}^0$ is the fluctuation of the bond connecting atoms i and j , where r_{ij}^0 is the equilibrium bond distance. The spring constants, k_{ij} , define the interactions between atoms i and j . There are different rules to determine force constants in ENM.^{47–57} Most popularly, a given cutoff distance is used to define the connections, and only the atoms within the cutoff distance are connected by springs, then a uniform force constant is placed for all connected atoms.^{47,49} Recently, Hinsen argued that distance-weighted interactions in ENM are physically better motivated, which are superior to the cutoff-based interactions in reproducing crystallographic B-factors.⁵⁴ Therefore a “parameter-free” ENM (pfENM) is used here, in which the force constants between atoms are weighted by the inverse square of their distances.⁵⁷

$$k_{ij} = c(r_{ij}^0)^{-2} \quad (2)$$

where c is a constant, which simply scales the overall range of B-factors.

For a system with n atoms, the second derivatives of the overall potential (eq 1) can be organized in a Hessian matrix $\mathbf{H} \in \mathbb{R}^{3n} \times \mathbb{R}^{3n}$. The elements of this matrix are

$$H(i_x, j_y) = \partial^2 V / \partial r_{i_x} \partial r_{j_y} \quad (3)$$

where r_{i_x} and r_{j_y} are the x ($= 1, 2, 3$) component of the position of the atom i , and the y ($= 1, 2, 3$) component of the position of the atom j , respectively. \mathbf{H} can be diagonalized to yield a matrix of eigenvectors and corresponding eigenvalues,

$$H(i_x, j_y) = \sum_{q=1}^{3n} \Psi_q^{i_x} \lambda_q \Psi_q^{j_y} \quad (4)$$

Here $\Psi_q^{i_x}$ and $\Psi_q^{j_y}$ are the two components corresponding to the x coordinate of the atom i and the y coordinate of the atom j , respectively, in the eigenvector $\Psi_q \in \mathbb{R}^{3n}$ (normal mode), which is the q^{th} column of the matrix $\Psi \in \mathbb{R}^{3n} \times \mathbb{R}^{3n}$. There are a total of $3n$ eigenvalues λ_q , the first six of which are zero because rigid-body translations and rotations leave the Hamiltonian invariant. Each nonzero eigenvalue and corresponding eigenvector represents the frequency and the Cartesian components of this normal mode, respectively. It should be noted that different choices of the constant c in eq 2 will only change the eigenvalues but not the eigenvectors. Many studies have indicated that the first few low-frequency normal modes describe functionally important motions in biomolecules.^{58,59}

ENM-ED-CG Method. The details of the ENM-ED-CG methodology are described in ref 21. From pfENM, an essential subspace is obtained, which consists of the first n_{ED} of the low-frequency normal modes with nonzero eigenvalues. For an N -site CG model of a biomolecule, $n_{\text{ED}} = 3N - 6$ since it has $3N - 6$ internal degrees of freedom. In the

biomolecule, those atoms that move together in a highly correlated fashion (called a dynamic domain) are mapped into one CG site by minimizing the following residual

$$\chi^2 = \frac{1}{3N} \sum_{I=1}^N \sum_{i \in I} \sum_{j \geq i \in I} \langle (\Delta \mathbf{r}_i^{\text{ED}})^2 - 2\Delta \mathbf{r}_i^{\text{ED}} \cdot \Delta \mathbf{r}_j^{\text{ED}} + (\Delta \mathbf{r}_j^{\text{ED}})^2 \rangle \quad (5)$$

where $\Delta \mathbf{r}_i^{\text{ED}}$ is the fluctuation of atom i in the essential subspace. If another atom j moves together with the atom i , their fluctuation difference, $|\Delta \mathbf{r}_i^{\text{ED}} - \Delta \mathbf{r}_j^{\text{ED}}|^2$, would be very small. Thus the residual (eq 5) can be minimized by grouping them into the same CG site I . A CG model defined by this algorithm can therefore preserve dynamic domains in the atomistic model and approximate the functional essential dynamics of the biomolecule.

According to the classical theory of networks,⁶⁰ the mean-square fluctuation of atom i in the essential subspace is

$$\langle (\Delta \mathbf{r}_i^{\text{ED}})^2 \rangle = k_B T \text{tr}[(\mathbf{h}^{\text{ED}})_{ii}^{-1}] \quad (6)$$

where k_B is the Boltzmann constant, and T is the absolute temperature. The term $(\mathbf{h}^{\text{ED}})_{ii}^{-1} \in \mathbb{R}^3 \times \mathbb{R}^3$ is the i^{th} diagonal superelement (a 3×3 matrix) in the inverse matrix of $\mathbf{H}^{\text{ED}} \in \mathbb{R}^{3n} \times \mathbb{R}^{3n}$ (\mathbf{H} in the essential subspace), and $\text{tr}[\]$ represents the trace. That is to say,

$$\text{tr}[(\mathbf{h}^{\text{ED}})_{ii}^{-1}] = \sum_{x=1}^3 \sum_{q=7}^{n_{\text{ED}}+6} \Psi_q^{i_x} \lambda_q^{-1} \Psi_q^{i_x}$$

It follows that eq 5 may be recast in the following form:

$$\chi^2 = \frac{k_B T}{3N} \sum_{I=1}^N \sum_{i \in I} \sum_{j \geq i \in I} (\text{tr}[(\mathbf{h}^{\text{ED}})_{ii}^{-1}] - 2\text{tr}[(\mathbf{h}^{\text{ED}})_{ij}^{-1}] + \text{tr}[(\mathbf{h}^{\text{ED}})_{jj}^{-1}]) \quad (7)$$

The original ENM-ED-CG method systematically defined CG sites in a protein along its primary amino acid sequence. However, the atomistic details and sequence information underlying a low-resolution structure may not be known, which precludes the coarse graining of many proteins and the protein complexes using the ENM-ED-CG scheme. A different approach is therefore needed and described in the next two sections.

Discretization of a Density Map. Low-resolution structural information, such as density maps measured by cryo-EM or SAXS, can be discretized using a technique called vector quantization (VQ).^{28–31} The VQ approach allows one to represent a density map by a finite number of n pseudoatoms, which approximate the density (mass distribution) according to a statistical optimization criterion. There are several algorithms and utilities to solve the VQ problem. In the program packages Situs^{61,62} and Sculptor,⁶³ a VQ tool is provided for generating a pseudoatom model given an input volumetric structure, by using a so-called “neural gas” network algorithm.^{30,31} This approach has also been implemented in the VMD program⁶⁴ recently. In the present work, Sculptor is used to generate a pseudoatom model from a density map.

A Space-Based ED-CG Scheme. After a density map is discretized into n pseudoatoms, a pFNM is built. Then one can define a CG model with N sites from the pseudoatom model. The spirit of ED-CG is to group atoms that move in a highly correlated fashion into a CG site. When the atoms are close in three-dimensional (3D) space, they may have a high tendency to move together and define the best CG unit in the ED-CG scheme. Therefore a space-based ED-CG algorithm is proposed for the pseudoatom model. The details of the algorithm are as follows:

- (1) N cluster seeds are generated randomly. In practice, the position of one seed S (R_{S_x} , $x = 1-3$) is the position of one randomly selected pseudoatom, plus a small random offset (from -1.0 to 1.0).
- (2) The pseudoatoms are clustered into N groups (domains) according to the N seeds, such that every atom in one domain is closer to the corresponding seed than any other $N - 1$ seeds. Once the N domains are determined, the position of central-of-mass (COM) of each domain, denoted as I , is computed (R_{I_x} , $x = 1-3$).
- (3) The average difference between R_{I_x} and R_{S_x} is calculated as

$$R_{\text{diff}} = \frac{1}{3N} \sum_{I=1}^N \sum_{x=1}^3 |R_{I_x} - R_{S_x}| \quad (8)$$

Sometimes R_{diff} is rather large, which may indicate that in one or more domains, the atoms are not close in space. In the worst case, one domain may contain separate pieces that are far apart. To avoid this, R_{diff} needs to be decreased in the following way. The positions of the cluster seeds are updated/replaced by the COM of the domains, that is $R_{S_x} = R_{I_x}$. Repeat step 2 with the new cluster seeds to get the updated positions of the COM of the domains and a new R_{diff} (eq 8). Repeat this until R_{diff} is below a certain value $R_{\text{diff}}^{\text{max}}$.

- (4) The CG sites are taken as the COM of the domains, and the residual of this N -site model is calculated by eq 7.
- (5) Randomly pick a cluster seed and update its position. Repeat the steps 2–4 and obtain an updated CG model with a new residual. This new CG model is accepted or rejected based on the Metropolis criterion,⁶⁵ as introduced in refs 20, 21. Step 5 is iterated for a number of steps (n_{SA}) to minimize the residual (eq 7) using a simulated annealing algorithm.⁶⁶
- (6) The above steps 1–5 are performed, beginning with different initial sets of cluster seeds. Finally, the CG model with the lowest residual is taken.

It should be noted that step 3 is used to avoid domains with distant groups of atoms, which cannot be achieved by minimizing the residual (eq 7) only. Even if two groups of atoms are far apart in space, their fluctuation difference could still be small, and thus the residual may be minimized if they are grouped into the same CG site. Step 3 serves as a constraint to find domains without distant groups of atoms while minimizing the residual (eq 7).

One might imagine that the value of $R_{\text{diff}}^{\text{max}}$, the maximum allowed difference between the COM of the domains and

the cluster seeds, is a critical parameter which determines the size of the space for ED-CG searching. When $R_{\text{diff}}^{\text{max}}$ is too small, the searchable space for ED-CG is highly limited, and the final CG model will mostly depend on eq 8 rather than eq 7. Actually, eq 8 is a so-called Linde–Buzo–Gray (LBG) algorithm to solve the VQ problem when $R_{\text{diff}}^{\text{max}}$ is very small.²⁸ A larger value of $R_{\text{diff}}^{\text{max}}$ can certainly broaden the searchable space for ED-CG but may render this constraint meaningless. We have tested different values and found a $R_{\text{diff}}^{\text{max}}$ around 0.5 Å to be a good compromise. The values within this range could avoid distant groups of atoms in the same domain, and in the mean time, a reasonable space for ED-CG searching was allowed. In this work, $R_{\text{diff}}^{\text{max}}$ is therefore chosen as 0.5 Å.

CG Models from Different Methods. For all the systems studied in the next section, the corresponding space-based ED-CG models will be mainly presented by using the new algorithm described above, which can be applied to both an atomistic structure and a pseudoatom model from a density map. CG models from other methods are also introduced for comparison. The VQ method can define space-based CG models, from the atomistic structure, as well as the pseudoatom model or the density map directly. The previous published ENM-ED-CG scheme²¹ can be used to define sequence-based ED-CG models only if the atomistic structure of the system is available. Therefore, three kinds of CG models (space- and sequence-based ED-CG and VQ-CG models, respectively) are all discussed when defining CG models from an atomistic structure of the system. For a density map, a pseudoatom model is constructed by the VQ method. In that case, both the space-based ED-CG and VQ-CG models are defined based on the pseudoatom model.

Results and Discussion

CG Models of G-Actin from Atomistic Structure. The protein G-actin is a globular protein with 375 residues,⁶⁷ which constitutes the subunit of the actin filament.^{68–70} Low-resolution CG models of the G-actin have been widely used to explore the elastic properties of the actin filament.^{71–73} The atomistic structure of the G-actin is available and was used in the previous two ED-CG papers^{20,21} to define sequence-based CG models. As a comparison, the space-based ED-CG method was also applied to the atomic model of the G-actin to define space-based ED-CG models. Only the 375 C $_{\alpha}$ atoms in the atomic structure were used to build a pfENM.

Four-Site CG Models. By usual inspection, G-actin (Figure 1) can be divided into four spatial domains.⁶⁸ The residue numbers of these domains are D1 (1–32, 70–144, and 338–375; Figure 1a, blue), D2 (33–69; Figure 1a, red), D3 (145–180 and 270–337; Figure 1a, orange), and D4 (181–269; Figure 1a, green). By taking the COM of each domain as a CG site, Chu and Voth^{71,72} have developed an intuitive four-site CG model of G-actin, which is in fact a space-based CG model (Figure 1a). The first mode from pfENM indicates a propeller-like motion of the domains, which is in agreement with the atomistic normal-mode analysis.⁷⁴ In this mode (Figure 1a), the domain D1 has a

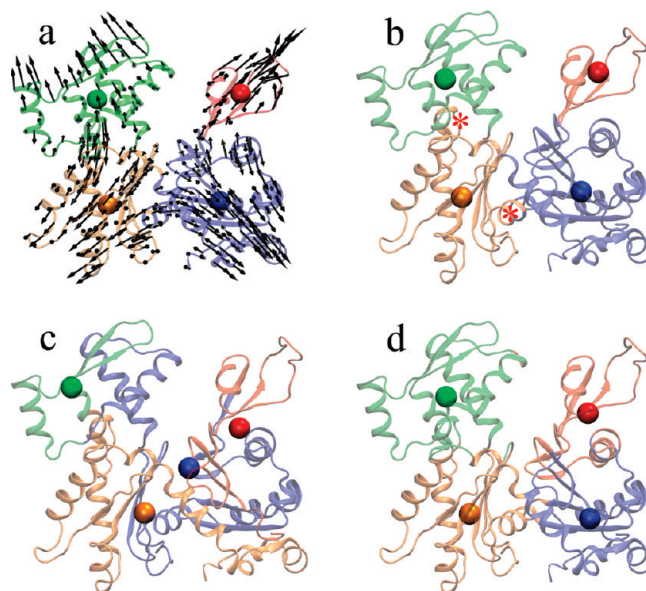


Figure 1. Four-site models of the G-actin. (a) The intuitive four-site model: D1 (1–32, 70–144, 338–375) blue; D2 (33–69) red; D3 (145–180, 270–337) orange; and D4 (181–269) green. First mode from pfENM is shown by arrows. (b) Space-based ENM-ED-CG four-site model: D1 (1–33, 70–140, 337–375) blue; D2 (34–69) red; D3 (141–181, 261–336) orange; and D4 (182–260) green. Its differences to the intuitive four-site model are marked by asterisks. (c) Sequence-based ENM-ED-CG four-site model: (1–66) red; (67–219) blue; (220–256) green; and (257–375) orange. (d) VQ four-site model. Figures 1 and 3–9 were created using VMD.⁶⁴

tendency of coming into the plane of the page as the domain D2 moves out of the plane and vice versa. The domains D3 and D4 perform a similar but antiparallel motion to the domains D1 and D2. That is to say, the domain D4 comes into the plane, while the domain D2 moves out of the plane and vice versa. The intuitive four-site model of G-actin thus naturally allows one to study this propeller motion, which may be related to the opening/closing of the ATP binding cleft.⁶⁷

Here, a four-site CG model was defined with the space-based ED-CG method, using the first six normal modes ($n_{\text{ED}} = 6$). Cluster seeds from 4000 random initial sets were used, and $n_{\text{SA}} = 5000$ steps were calculated for each set, to minimize the residual (eq 7). The lowest residual after these SA iterations was 5391 but only two out of the total 4000 cluster-seed sets reached the lowest residual value. The result indicates that the convergence of the space-based ED-CG method is not as good as the previous sequence-based method.^{20,21} The space-based search is much more complicated than the sequence-based one, and it would be difficult to sufficiently sample all the possibilities in limited steps. Nevertheless, the space-based ED-CG four-site model with the lowest residual after the SA iterations looks similar to the intuitive four-site model, but the latter has a higher residual of 5458.

There are seven sequence-contiguous subdomains in the intuitive four-site model (Figure 2a), and the space-based ED-CG model has seven similar subdomains as well,

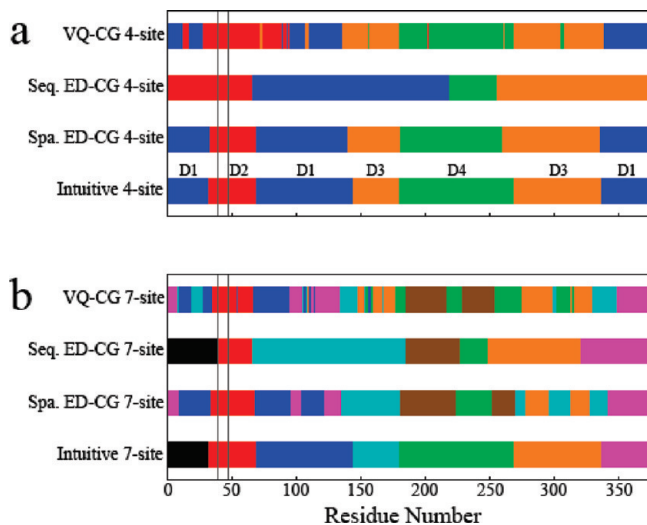


Figure 2. Allocation of domains in different CG models of the G-actin. (a) Four- and (b) seven-site models. The x axis is the residue number of the G-actin, and y axis represents different CG models. Four subdomains (D1–D4) are indicated on top of the intuitive four-site model. Domains are colored according to Figures 1 and 3 (four- and seven-site models, respectively). DB-loop region (residue 40–48) is marked with vertical gray lines.

although they are not completely contiguous in sequence. For example, there is one subdomain that is contiguous from residues 70–144 in the intuitive model, and the corresponding subdomain in the space-based ED-CG model spans the residues 70–140. Most of the residues from 70–140 belong to the same subdomain but a few interspersed residues belong to other subdomains. Interestingly, if we smoothed the subdomain by changing these a few residues to the same subdomain that most residues (from 70–140) belong to, the residual was further minimized.

It was relatively straightforward to smooth the space-based ED-CG model obtained by the SA iterations along the primary sequence, in order to make the contiguous subdomains while minimizing the residual further. Thus a new space-based ED-CG four-site model was obtained with a little lower residual of 5335 (Figure 1b) than the one without smoothing along the sequence. The residue numbers of the domains in this model are D1 (1–33, 70–140, and 337–375; Figure 1b, blue), D2 (34–69; Figure 1b, red), D3 (141–181, and 261–336; Figure 1b, orange), and D4 (182–260; Figure 1b, green). This space-based ED-CG four-site model is very close to the intuitive four-site model (Figure 1a) with just a few differences (marked by asterisks in Figure 1b). For example, the domain D4 is from residues 181–269 in the intuitive four-site model, but it is from residues 182–260 in the space-based ED-CG four-site model. According to the motions of residues 261–269 in the first ENM mode, it is better to place them in the domain D3 since they move correlated with the other residues in this domain (Figure 1b). That is to say, the space-based ED-CG four-site model is somewhat better at dividing the dynamics domains than the intuitive one. These results indicate that the space-based ED-CG method can define a robust CG model that is consistent with intuition but in a more systematic and quantitative way.

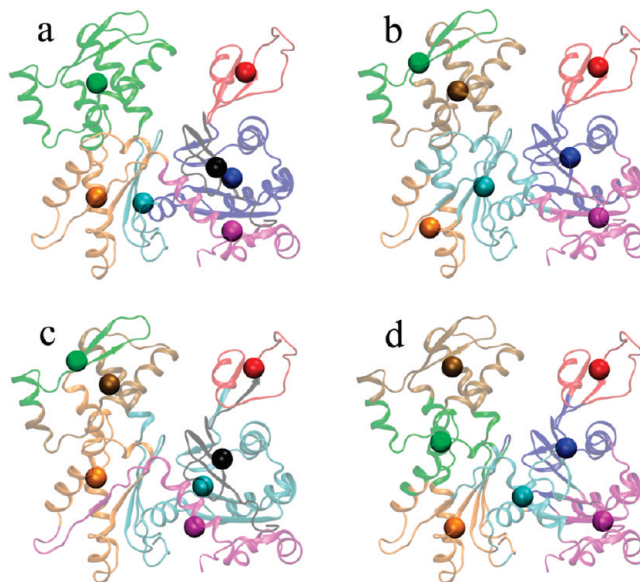


Figure 3. Seven-site models of the G-actin. (a) The intuitive seven-site model that is a sequence-based model: (1–32) black; (33–69) red; (70–144) blue; (145–180) cyan; (181–269) green; (270–337) orange; and (338–375) magenta. (b) The space-based ENM-ED-CG seven-site model: (1–39) black; (40–66) red; (67–185) cyan; (186–227) ocher; (228–249) green; (250–321) orange; and (322–375) magenta. (c) The sequence-based ENM-ED-CG seven-site model: (1–39) black; (40–66) red; (67–185) cyan; (186–227) ocher; (228–249) green; (250–321) orange; and (322–375) magenta. (d) VQ seven-site model.

A sequence-based four-site model (Figure 1c) was also defined by the previous ENM-ED-CG method.²¹ A total of 86 out of the total 200 initial boundary atom sets reached the minimal residual of 5672, which indicated a better convergence in the sequence-based ED-CG search than the space-based one. The model contains four sequence-contiguous domains: (1–66; Figure 1c, red), (67–219; Figure 1c, blue), (220–256; Figure 1c, green), and (257–375; Figure 1c, orange). These domains, which obviously deviate from intuition, are very different from those in the space-based ED-CG four-site model (Figure 1b) because of the additional constraint of having primary sequence-contiguous domains. Furthermore, the domain motions in the low-frequency normal mode (Figure 1a) cannot be described properly by the sequence-based ED-CG four-site model, which explains why it has a significantly higher residual than the space-based ED-CG four-site model. That is, the sequence-based model is not as good at preserving the essential low-frequency dynamics of G-actin as the space-based ED-CG model for this highly coarse-grained (four sites) model.

A four-site model defined by the VQ-CG method is shown in Figure 1d. This model looks like the space-based ED-CG and intuitive models (Figure 1b and 1a, respectively) but with a significantly higher residual of 5865. Instead of the seven sequence-contiguous subdomains in those two models, the VQ-CG four-site model becomes rather mixed along the primary sequence (Figure 2). This result is actually quite important because it shows that having any motion of the protein included in the CG model development (even an ENM motion based on VQ pseudoatoms) causes the resulting CG model to have a much greater (and physical) “molecule-like” character.

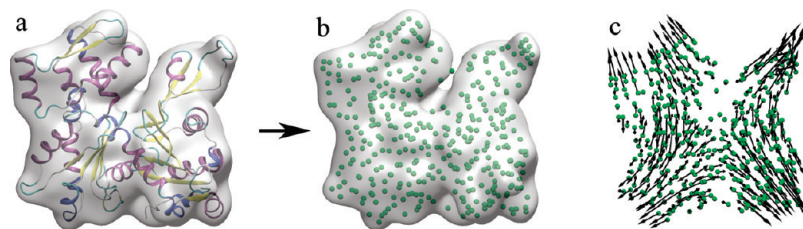


Figure 4. (a) From the atomistic structure of the G-actin, a density map of 10 Å resolution was generated by Situs.^{61,62} Then a (b) pseudoatom model with 375 pseudoatoms was built by the VQ algorithm in Sculptor.⁶³ (c) First mode from pfENM of the pseudoatom model.

Seven-Site CG Models. The intuitive four-site model of G-actin consists of seven sequence-contiguous subdomains (Figures 1a and 2a) from which an intuitive seven-site model was defined (Figure 2b and Figure 3a). It is a sequence-based model with a high residual of 4334, where the first 15 normal modes were used ($n_{ED} = 15$).

A space-based ED-CG seven-site model was defined by using 4000 random initial sets of cluster seeds and 5000 SA steps for each set. The model was then smoothed along the primary sequence to further minimize the residual, and the final seven-site model is shown in Figure 3b. This model has a much lower residual (2410) than the intuitive seven-site model (Figure 3a), which means it does a much better job of preserving the essential protein dynamics. The space-based ED-CG seven-site model obviously includes more detail than the space-based ED-CG four-site model (Figure 1b). The domain D2, which includes the most flexible DB-loop,⁷⁰ almost remains the same between the two models (Figures 1b and 3b, red). The other three sites (blue, green, and orange) in the space-based ED-CG four-site model (Figure 1b) are all divided into two CG sites, respectively, in the space-based ED-CG seven-site model (Figure 3b). Besides the DB-loop, residues 220–252 in the D4 domain are also highly mobile according to the first ENM mode (Figure 1a). Therefore it is defined as a separate CG site in the space-based ED-CG seven-site model (Figure 3b), in order to minimize the residual. However in the intuitive seven-site model (Figure 3a), the domain D4 remains intact as that in the intuitive four-site model (Figure 1a). Instead the domain D1, which is less flexible, is divided into three domains (black, blue, and magenta in Figure 3a). That explains why the residual of the intuitive seven-site model is so high because it does not capture these highly mobile domains properly.

The sequence-based ED-CG seven-site model (Figure 3c) has a residual of 2701, which is higher than the space-based ED-CG seven-site model (Figure 3b) but still much lower than the intuitive seven-site model (Figure 3a). Three CG sites (red, green, and ocher sites in Figure 3c) are similar to those in the space-based ED-CG seven-site model (Figure 3b). Therefore the residual of the sequence-based ED-CG seven-site model is reasonably low, since it can well describe these mobile regions in the domains D2 and D4.

Interestingly the residual of the VQ-CG seven-site model (Figure 3d) is 3185, still significantly lower than the intuitive seven-site model (Figure 3a). In fact, the sites in the VQ-CG seven-site model (Figure 3d) look like those in the space-based ED-CG seven-site model (Figure 3b), except for the

relative locations between the green and ocher sites. However the VQ-CG seven-site model significantly scrambles the primary protein sequence (Figure 2b), making it wonder how the motions of such a CG model would correspond to underlying atomistic motions.

CG Models of G-Actin from Density Map. By lowering the resolution of the atomistic structure of G-actin, one can create a density map at a specified resolution (Figure 4a). This was done with the Situs package^{61,62} by using its program “pdb2vol”. A volumetric density map of G-actin at 10 Å resolution was thus generated (Figure 4a). Then 375 pseudoatoms (Figure 4b) were defined to represent the density map by using the VQ method in Sculptor,⁶³ which are not the same as the 375 C α atoms from the high-resolution atomic structure. However, the first mode from pfENM of the pseudoatom model exhibits a similar propeller-like motion (Figure 4c) as in the first mode from the atomistic structure (Figure 1a). The space-based ED-CG method was then applied to group the pseudoatoms into CG sites. As in the case of the atomic structure, the same number (4000) of random initial sets of cluster seeds and SA steps (5000) were used. For a single initial cluster-seed set, the calculation was finished in a few seconds on a 2.4 GHz desktop personal computer.

Four-Site CG Models. The space-based ED-CG four-site model from the pseudoatom model has the lowest residual of 5127 (it should be noted here that residuals between the pseudoatom and atomistic models are not comparable). This CG model (Figure 5a) is very similar to that from the atomic model (Figure 1b). A VQ-CG four-site model (Figure 5b) from the same pseudoatom model is also rather close to the space-based ED-CG four-site model (Figure 5a), which is consistent with the results from the atomistic structure of G-actin (Figure 1b and d). However the VQ-CG four-site model has a higher residual of 5352 than the space-based ED-CG four-site model.

Seven-Site CG Models. The space-based ED-CG seven-site model, with the lowest residual of 2213, is shown in Figure 5c. The model looks similar to the space-based ED-CG seven-site model from the atomistic structure (Figure 3b), except that the relative positions of the green and ocher sites and the cyan and orange sites between them are somewhat different. The red CG site that includes the critically important DB-loop⁷⁰ appears the same in both the ED-CG four- and seven-site models from the pseudoatom model. The other three CG sites (blue, green, and orange) in the space-based ED-CG four-site model (Figure 5b) are

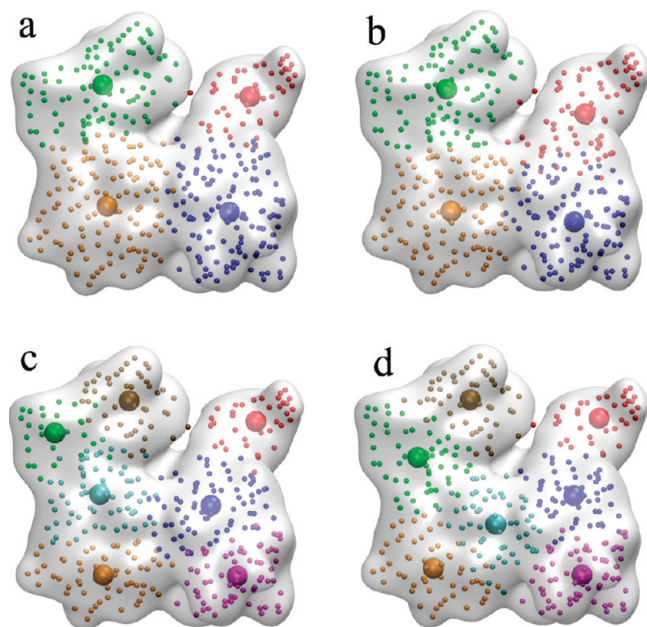


Figure 5. Space-based CG models of the G-actin from its pseudoatom model (Figure 4c). (a) Space-based ED-CG four-site model. (b) VQ-CG four-site model. (c) Space-based ED-CG seven-site model. (d) VQ-CG seven-site model.

divided into two sites each in the space-based ED-CG seven-site model (Figure 5c). These results are consistent with those derived from the underlying atomic structure of G-actin (Figures 1b and 3b), which suggests that the space-based ED-CG method also does a good job of preserving dynamic domains derived from the pseudoatom model. The VQ-CG seven-site model from the pseudoatom model (Figure 5d) may not be adequate in this regard because some of its CG sites, such as the green and cyan sites, cross different domains in the four-site model (Figure 5a). That may explain why the VQ-CG seven-site model has a higher residual (2623) than the space-based ED-CG seven-site model.

The above results are quite encouraging, demonstrating that it is a viable strategy to discretize a low-resolution density map into pseudoatoms by VQ and then to apply the space-based ED-CG approach directly to the pseudoatom model to define CG sites that preserve the essential dynamics of the system. The space-based ED-CG models built from these pseudoatoms are quite similar to those built from an underlying high-resolution atomistic structure directly. In the following sections, the space-based ED-CG method will therefore be applied directly to density maps determined by cryo-EM.

CG Models of *E. coli* 70s Ribosome from Density Map. The *E. coli* 70S ribosome is a large RNA–protein complex, which plays a central role in protein biosynthesis.^{1–3} The complete ribosome consists of a small and large subunit. The 30S small subunit contains a 16S rRNA and about 20 S proteins. The 50S large subunit contains a 23S and 5S rRNA and over 30 L proteins. To date, bacterial ribosome structures are available from both low-resolution cry-EM density maps^{75,76} and high-resolution atomic structures.^{77–82} It is computationally very expensive to study the functional dynamics of the ribosome by atomistic MD simulations since it is a very large macromolecular assembly.^{83–86} Coarse-

grained ENMs have predicted certain global motions in the ribosome, such as the ratchet-like reorganization between the small and large subunits.^{43–46} In this work, a cryo-EM density map of the ribosome (Figure 6a) at a 11.5 Å resolution⁷⁶ is used to define space-based ED-CG models. A total of 2000 pseudoatoms were generated by the VQ method (Figure 6b) from the ribosome density map, and a pfENM was built from the pseudoatom model. The first normal mode (Figure 6c) does describe a ratchet-like motion between the small and large subunits, which is in agreement with the results from experiments^{75,87} and other atom-based ENMs.^{43–46} We then define space-based ED-CG models directly from this pseudoatom model.

40-Site CG Ribosome Models. To define a space-based ED-CG 40-site model, 2000 random initial sets of cluster seeds were used, and $n_{SA} = 80\,000$ steps were performed for each set to minimize the residual (eq 7). For one initial set of cluster seeds, the SA minimization was done in about 4–5 min on a 2.4 GHz CPU. To speed the calculations of the 2000 initial sets, they were distributed onto multiple computer nodes at the same time, since they were completely independent. The model with the lowest residual (535) is shown in Figure 7a. For comparison, a VQ-CG 40-site model was also generated (Figure 7b) that has a much higher residual of 933. Although the resolution of the ribosome density map is only 11.5 Å, the small and large subunits and some other structural details (like the head, spur, L7/L12 stalk base, and protein L1) are visible (Figure 6a). In the space-based ED-CG 40-site model, there are 17 (Figure 7a, orange) and 21 sites (Figure 7a, blue) in the small and large subunits, respectively, and 2 sites are located in the bridges between the two subunits (Figure 7a, magenta). In the VQ-CG 40-site model, there are 13 (Figure 7b, orange) and 25 sites (Figure 7b, blue) in the small and large subunits, respectively, and 2 sites in the bridge area (Figure 7b, magenta).

Since the molecular weight of the small subunit is approximately half of the weight of the large subunit,¹ the CG-site distribution in the VQ-CG 40-site model reflects the mass distribution in the ribosome. However, the space-based ED-CG 40-site model contains more CG sites in the small subunit than in the VQ-CG 40-site model. It is well-known that the 30S small subunit fluctuates more than the 50S subunit in the ribosome dynamics.^{75,77,78} Therefore, more CG sites are located in the small subunit relative to its size, in order to better represent its dynamics. For example, the spur region in the small subunit has large fluctuations, according to the first normal mode (Figure 6c). The space-based ED-CG 40-site model has 2 sites in this region to preserve its dynamics (Figure 7a), but the VQ-CG 40-site model only defines 1 site in that region (Figure 7b). The site that represents the spur region in the VQ-CG 40-site model has a too large fluctuation (a so-called “tip effect”),⁵⁰ which means a single site may be not enough to well describe the dynamics of the spur.

The same behavior happens in the 50S large subunit. Although the space-based ED-CG 40-site model has less CG sites in the large subunit than the VQ-CG 40-site model, more CG sites are located in the functionally important regions, such as the L1 and L7/L12 stalks (Figure 7a). The

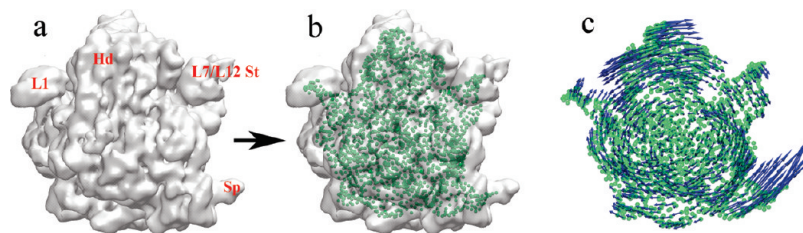


Figure 6. From (a) the 11.5 Å density map of *E. coli* 70S ribosome,⁷⁶ a (b) pseudoatom model with 2000 pseudoatoms was built by the VQ algorithm in Sculptor.⁶³ (c) First normal mode from pfENM of the pseudoatom model, which describes the ratchet-like motion between the small and large subunits.

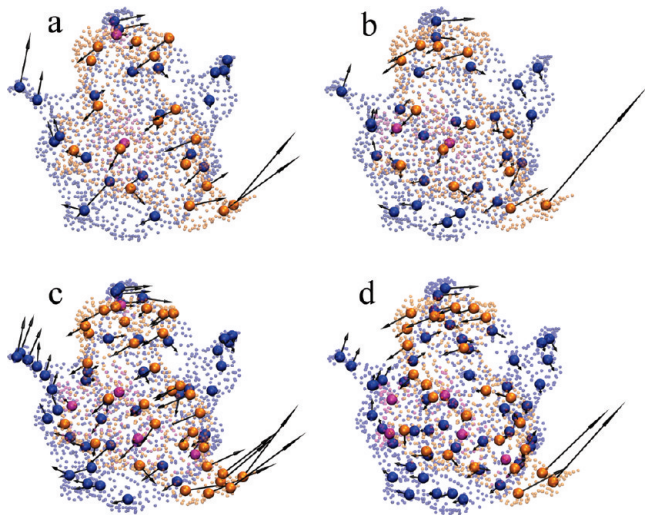


Figure 7. Space-based CG models of the ribosome from its pseudoatom model (Figure 6b). (a) Space-based ED-CG 40-site model. (b) VQ-CG 40-site model. (c) Space-based ED-CG 80-site model. (d) VQ-CG 80-site model. CG sites that belong to the small subunit are colored by orange, CG sites that belong to the large subunit are colored by blue, and those CG sites that make the bridges between the subunits are colored by magenta. In each CG model, the first normal mode from pfENM is shown by arrows.

body of the 50S subunit is massive, so a large number of CG sites are needed in the VQ-CG 40-site model in order to reflect its mass distribution (Figure 7b). However the 50S body is fairly rigid, and fewer CG sites are sufficient to preserve its dynamics (Figure 7a). According to experimental data,⁷⁶ there are some intersubunit bridges that are located in both the head and body regions of the ribosome. For the two bridge CG sites in the space-based ED-CG 40-site model (magenta in Figure 7a), one site is located in the head, and the other one is in the body of the ribosome. However in the VQ-CG 40-site model, both bridge CG sites (magenta in Figure 7b) are in the body because it is much more massive than the head. The locations of the bridge sites in the space-based ED-CG 40-site model better describe the association between the two subunits. In summary, the space-based ED-CG model appears to be superior in identifying and preserving the functional dynamics between the ribosome subunits.

80-Site CG Ribosome Models. To define a space-based ED-CG 80-site model from the pseudoatom model of the ribosome, again 2000 initial sets of cluster seeds were used and $n_{SA} = 160\,000$ steps were performed for each set to

minimize the residual (eq 7). It took about 7–8 min to minimize a single cluster-seed set. The lowest residual of the space-based ED-CG 80-site model (Figure 7c) is 221, while the VQ-CG 80-site model (Figure 7d) has a higher residual of 376. Although 80 CG sites are still very coarse compared to the size of the ribosome, the space-based ED-CG 80-site model adds more detail than the space-based ED-CG 40-site model. As with the 40-site models, the regions with functional importance in the ribosome are better represented in the space-based ED-CG 80-site model than in the VQ-CG 80-site model. The small subunit contains 32 CG sites in the space-based ED-CG 80-site model (Figure 7c, orange), but this number in the VQ-CG 80-site model is only 26 (Figure 7d, orange). Five CG sites are defined to represent the dynamics of the spur region in the space-based ED-CG 80-site model (Figure 7c, orange), whereas the VQ-CG 80-site model has only 2 sites in this region (Figure 7d, orange). The space-based ED-CG 80-site model has a smaller number of 43 CG sites in the large subunit (Figure 7c, blue) than that of the VQ-CG 80-site model (49 sites, Figure 7d, blue). Nevertheless the former has more CG sites located in the L1 region and the L7/L12 stalk than the latter. There are five CG sites defined for the bridges in the space-based ED-CG 80-site model, one is in the head and the other four are in the body (magenta in Figure 7c). Importantly, the locations of these intersubunit bridges are in agreement with the experimental data.⁷⁶ In the VQ-CG 80-site model, all the five bridge sites are in the body (magenta in Figure 7d).

CG Models of the Microtubule from Density Map. Microtubules (MTs) are long and stiff hollow cylindrical tubes in eukaryotic cells, which play fundamental roles in many cellular processes, such as mitosis, cytokinesis, and vesicular transport.^{6–11} The structural subunit of a MT is the $\alpha\beta$ -tubulin heterodimer.^{88–93} The MT assembly involves two steps: the tubulin dimers bind head to tail to form protofilaments (pfs), and then pfs assemble side by side to complete the microtubule.⁹⁴ The atomic structure of the tubulin dimer has been determined by electron crystallography⁸⁹ and refined to 3.5 Å resolution.⁹² However, MTs have not yet been found to be suitable for X-ray crystallography, since they are highly polymorphic. In this case, cryo-EM was well adapted for obtaining a 3D reconstruction of the MT.^{95–98}

In this work, space-based ED-CG models are defined from an EM density map of the MT at an 8 Å resolution.⁹⁸ In this 3D map (Figure 8a), there are 13 parallel pfs, and lateral interactions between them complete the MT wall. In each pf, there are approximately four tubulin monomers (three

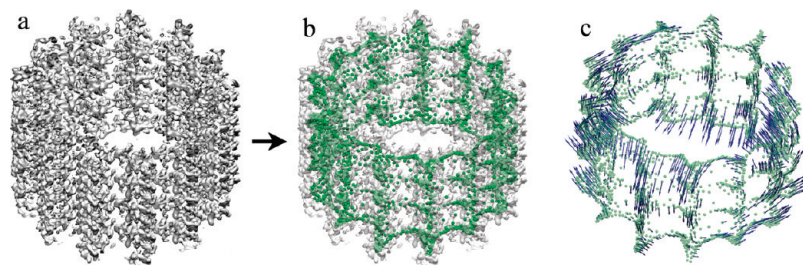


Figure 8. From (a) the 8.0 Å density map of the microtubule,⁹⁸ with its plus end towards the top, (b) pseudoatom model with 3200 pseudoatoms was built by the VQ algorithm in Sculptor.⁶³ (c) First mode from pfENM of the pseudoatom model.

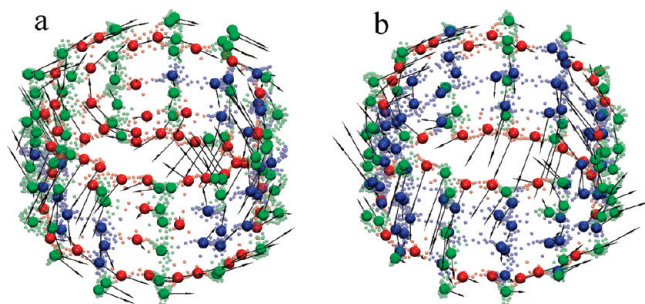


Figure 9. Space-based CG models of the microtubule from its pseudoatom model (Figure 8b). (a) Space-based ED-CG 156-site model. (b) VQ-CG 156-site model. Longitudinal CG sites that represent the pseudoatoms along the pfs are colored by green, lateral CG sites that represent the pseudoatoms between the pfs are colored by red, and mixed CG sites that contain pseudoatoms both along and between the pfs are colored by blue. First pfENM mode of each model is shown by arrows, respectively.

complete monomers in the middle, an incomplete monomer on the top, and another incomplete one at the bottom), which are connected though longitudinal contacts. At this resolution, the α and β tubulins are essentially not distinguishable because their structures are very similar.⁹⁹ However, many secondary structures are still visible in the monomers.

156-Site CG Microtubule Models. A total of 3200 pseudoatoms were generated by the VQ method (Figure 8b) from the MT density map, and the first pfENM mode of this pseudoatom model is shown in Figure 8c. As far as we know, no work about the functional modes of the MT has been reported yet, except for the tubulin dimer.^{100,101} The top mode of the MT (Figure 8c) indicates a bending motion of the MT and a twist between pfs, which may change the shape of the MT. The fluctuations of the plus and minus ends are antiparallel. A space-based ED-CG 156-site model was directly defined from the pseudoatom model. A total of 2240 initial sets of cluster seeds were used, and $n_{SA} = 500\,000$ steps were performed for each set to minimize the residual (eq 7). This system is larger than the ribosome, and more CG sites were defined, therefore, almost 1 h was needed to complete the calculation of a single set of cluster seeds. The space-based ED-CG 156-site model (Figure 9a) has the lowest residual of 264, whereas the VQ-CG 156-site model (Figure 9b) has a higher residual of 366. In the space-based ED-CG 156-site model, the top and bottom of the MT (incomplete monomers) have more CG sites than the complete monomers in the middle (Figure 9a). The pseudoatoms near the top and bottom are more flexible in pfENM

due to fewer interactions than those in the middle (Figure 8b); therefore, more CG sites reside in the top and bottom regions of the MT in order to represent their larger fluctuations (Figure 8c). In the VQ-CG 156-site model, the distribution of CG sites between the top, bottom, and middle is more even (Figure 9b).

Another notable feature of the space-based ED-CG 156-site model is that more CG sites are found to represent lateral interactions between pfs (Figure 9a). In the MT, longitudinal contacts between tubulin monomers are stronger than lateral contacts between pfs.^{94,102} By looking at the pseudoatom model (Figure 8b), many pseudoatoms are located along the pfs with a high density, but the pseudoatoms between the pfs are much less. According to the representing pseudoatoms, the CG sites are divided into three groups: CG sites that represent the pseudoatoms along the pfs (longitudinal sites, Figure 9, green), CG sites that represent the pseudoatoms between the pfs (lateral sites, Figure 9, red), and CG sites that contain pseudoatoms both along and between the pfs (mixed sites, Figure 9, blue). In the space-based ED-CG 156-site model, there are 74 longitudinal, 55 lateral, and 27 mixed sites. The corresponding numbers in the VQ-CG 156-site model are 59, 34, and 63, respectively. There are only 34 lateral sites in the VQ-CG 156-site model, which is consistent with its mass distribution. The majority of the CG sites (63) are mixed sites. It is believed that the lateral interactions between pfs are critical to regulate assembly/disassembly (dynamic instability)¹⁰³ of the MT,^{94,98} so fewer lateral CG sites indicates that the VQ-CG 156-site model may not be able to describe the MT dynamics as well. In contrast, there are a larger number (55) of the lateral CG sites in the space-based ED-CG 156-site model and only 27 mixed sites (Figure 9a). This distribution of CG sites suggests that the longitudinal and lateral interactions are more clearly separated in the space-based ED-CG model. The lateral interactions are better represented than those in the VQ-CG model by comparing the first mode between the two CG models (Figure 9a and b). The space-based ED-CG 156-site model seems better to preserve the essential MT dynamics than the VQ-CG 156-site model, so the former may be used to more faithfully study elastic properties of the MT at the CG level.¹⁰⁴

Conclusions

The sheer size of many large biomolecular complexes greatly complicates attempts to solve their high-resolution atomistic structures by X-ray crystallography or NMR. Alternative

techniques, such as cryo-electron microscopy (cryo-EM)^{23,24} and small-angle X-ray scattering (SAXS),^{25–27} provide low-resolution structural information in many cases. These large biological assemblies are therefore also ideal candidates for coarse-grained computational modeling. It is thus an important priority to directly define coarse-grained (CG) models using the available low-resolution structural data for these systems. The main focus of this article is a new and important extension of the essential dynamics coarse-graining (ED-CG) methodology,^{20,21} in order to build CG models directly from three-dimensional (3D) density maps obtained from cryo-EM or SAXS. First, a density map is discretized into a pseudoatom model by the vector quantization (VQ) method to retain the molecular shape,^{32–34} and a pfENM is then constructed. A number of studies suggests that such a pseudoatom model is indeed sufficient to describe the low-frequency dynamics of the biomolecular system because they are mainly shape dependent.^{105,106} Second, the essential dynamics defined by the low-frequency modes are used to determine space-based CG sites (eq 7). By definition in the present method, a CG site is the central-of-mass (COM) of a group of pseudoatoms that are close in space and move in a correlated way. Therefore, the search algorithm here is different from that in the previous sequence-based ED-CG method.^{20,21}

The space-based ED-CG method can certainly also be applied to detailed atomistic structures. The resulting space-based ED-CG four-site model from the G-actin atomistic structure (Figure 1b) is almost the same as the intuitive four-site model (Figure 1a) but has a slightly lower variational residual. Upon comparison with the sequence-based ED-CG four-site model (Figure 1c), the space-based ED-CG model looks much more reasonable, and its residual is significantly lower. For the seven-site CG models from the atomistic structure of G-actin, the space-based ED-CG model (Figure 3b) also has a significantly lower residual than the sequence-based model (Figure 3c). These results indicate that the space-based ED-CG algorithm may find a CG model which can be better suited to preserve the essential dynamics than the CG model found by using the sequence-based ED-CG method as long as no large-scale conformational changes (e.g., unfolding) occur between groups of atoms that are space-closed in the structure used for coarse-graining. However, the space-based ED-CG search is more complicated than the sequence-based search since the former needs to explore many more possibilities. Therefore, a global minimum could not be reached even for the four-site CG model of G-actin, and the residual must be further minimized by smoothing the domains along the protein sequence. The space-based ED-CG method is also computationally more expensive than the sequence-based ED-CG method because there are additional distance calculations when clustering the atoms (steps 2 and 3 in the space-based ED-CG algorithm).

For a biomolecule with an atomic-resolution structure, one can use either space- or sequence-based ED-CG methods to define CG models. If a CG site needs to represent a large number of atoms (a relative low-resolution CG model), such as the four-site model of G-actin, then the sequence-based model may be unreasonable (Figure 1c), and one should

instead choose the space-based ED-CG method (Figure 1b). For a relatively high-resolution model, i.e., each CG site represents only a small number of atoms, the sequence-based ED-CG method should perform well. In particular, when the system is large and the resolution of the CG model is high (i.e., many CG sites), the space-based ED-CG calculation becomes time consuming, and the sequence-based method is recommended.

For a density map with no atomic detail, the space-based ED-CG method is the only option to define ED-CG models. When a 10 Å density map of G-actin was created from its atomistic structure (Figure 4), the space-based ED-CG four-site model from the density map (Figure 5a) is found to be very close to the one obtained from the atomistic structure (Figure 1b). Furthermore, the space-based ED-CG seven-site model from the density map (Figure 5c) looks similar to the one from the atomistic structure (Figure 3b). These results indicate that the space-based ED-CG models from the low-resolution density map still do a good job of preserving the functional essential dynamics, which is in turn better than the corresponding VQ-CG models (Figure 5 b and d).

The ED-CG calculations from the density maps of the *E. coli* 70S ribosome and the microtubule (MT) (11.5 and 8 Å resolution, respectively) are very promising. At a very aggressive CG level, such as a 40-site representation of the ribosome, a space-based ED-CG model (Figure 7a) can still describe the ratchet-like motion between the small and large subunit and does so better than the VQ-CG 40-site model. Furthermore, regions that are important in functional dynamics, such as the head and spur in the small subunit and the L1 and L7/L12 stalks in the large subunit, contain more CG sites in the space-based ED-CG model than in the corresponding VQ-CG model (Figure 7b). In the space-based ED-CG 156-site model of the MT (Figure 9a), the lateral interactions between pfs are better represented than those in the VQ-CG 156-site model (Figure 9b). Therefore the space-based ED-CG model may be superior in preserving the MT dynamics (a precursor to dynamic instability) at the CG level. In the MT density map (Figure 8a), there are incomplete tubulin monomers on the top and at the bottom. After obtaining pfENM modes from the pseudoatom model, we can just use the subset of pseudoatoms that are in the two layers of complete monomers to define CG sites. A CG model of a longer MT may be built by duplicating the CG sites in this very short MT segment.

Computational cost of the space-based ED-CG method depends on both the number of pseudoatoms (n) from the density map and the number of CG sites (N) to be defined. The number of SA steps starting from one cluster-seed set is at least set as $n \times N$, and also the CPU time needed in one step is increasing with the system size. For a small system like G-actin, the calculation of a single set of cluster seeds is really fast (in a few seconds), and all the 4000 sets can be done within a couple of hours on a regular desktop computer. When systems become larger, such as the ribosome and the microtubule, they are computationally more expensive. However, the CPU time for a single cluster-seed set of the ribosome and the microtubule is still within 10

min and 1 h, respectively. Furthermore, all the cluster-seed sets can be distributed onto multiple CPUs (as many as one can have) at the same time, since they are independent of each other.

It should be noted that a space-based ED-CG N -site model is defined to capture the functional essential dynamics of a pseudoatom model constructed from a continuous density map by the VQ method. The number of pseudoatoms n should be much larger than the number of the CG sites N , since the ED-CG method is best suited to define a relatively small number of CG sites for a large biomolecule. One can also define a shape-based VQ-CG N -site model from the pseudoatom model, but it is not as good at describing the CG functional dynamics of the system as the space-based ED-CG model.

Before applying the space-based ED-CG method to general low-resolution structures, one should pay attention to the quality of density maps because they may contain noisy data. A real density map does not look like that shown in Figure 6a. There are actually many small satellite densities floating around the molecular surface. Fortunately these satellite densities can be ignored by using a proper threshold in the VQ calculation, so no pseudoatoms will be placed for these small noisy regions. Another subject of future research is to improve the convergence of the space-based ED-CG search. However, this approach as it stands now can lead the way to the systematic development of highly coarse-grained models of many large biomolecular complexes and thus ultimately to the CG computational modeling of such systems without the existence of high-resolution experimental structures.

Acknowledgment. This work is supported by a Collaborative Research in Chemistry grant from the National Science Foundation (CHE-0628257). The authors wish to thank Dr. Kenneth Downing for providing them with the 8 Å density map of the microtubule.⁹⁸ Z.Z. thanks Dr. Andrea Grafmüller, Dr. Edward Lyman, and Marissa Saunders for comments on the manuscript. Computer software is available upon request.

References

- (1) Steitz, T. A. *Nat. Rev. Mol. Cell Biol.* **2008**, *9*, 242–253.
- (2) Schmeing, T. M.; Ramakrishnan, V. *Nature* **2009**, *461*, 1234–1242.
- (3) Yonath, A. *J. R. Soc., Interface* **2009**, *6*, S575–S585.
- (4) Reisler, E.; Egelman, E. H. *J. Biol. Chem.* **2007**, *282*, 36133–36137.
- (5) Pollard, T. D.; Cooper, J. A. *Science* **2009**, *326*, 1208–1212.
- (6) Nogales, E. *Cell. Mol. Life Sci.* **1999**, *56*, 133–142.
- (7) Nogales, E. *Annu. Rev. Biochem.* **2000**, *69*, 277–302.
- (8) Nogales, E.; Wang, H. W.; Niederstrasser, H. *Curr. Opin. Struct. Biol.* **2003**, *13*, 256–261.
- (9) Nogales, E.; Wang, H. W. *Curr. Opin. Struct. Biol.* **2006**, *16*, 221–229.
- (10) Brun, L.; Rupp, B.; Ward, J. J.; Nedelec, F. *Proc. Natl Acad. Sci. U.S.A.* **2009**, *106*, 21173–21178.
- (11) van der Vaart, B.; Akhmanova, A.; Straube, A. *Biochem. Soc. Trans.* **2009**, *37*, 1007–1013.
- (12) Li, J.; Lykotrafitis, G.; Dao, M.; Suresh, S. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 4937–4942.
- (13) Karplus, M.; McCammon, J. A. *Nat. Struct. Biol.* **2002**, *9*, 646–652.
- (14) Adcock, S. A.; McCammon, J. A. *Chem. Rev.* **2006**, *106*, 1589–1615.
- (15) Tozzini, V. *Curr. Opin. Struct. Biol.* **2005**, *15*, 144–150.
- (16) Ayton, G. S.; Noid, W. G.; Voth, G. A. *Curr. Opin. Struct. Biol.* **2007**, *17*, 192–198.
- (17) Sherwood, P.; Brooks, B. R.; Sansom, M. S. P. *Curr. Opin. Struct. Biol.* **2008**, *18*, 630–640.
- (18) *Coarse-graining of condensed phase and biomolecular systems*; Voth, G. A., Ed.; CRC Press: New York, 2009.
- (19) Murtola, T.; Bunker, A.; Vattulainen, I.; Deserno, M.; Karttunen, M. *Phys. Chem. Chem. Phys.* **2009**, *11*, 1869–1892.
- (20) Zhang, Z.; Lu, L.; Noid, W. G.; Krishna, V.; Pfendner, J.; Voth, G. A. *Biophys. J.* **2008**, *95*, 5073–5083.
- (21) Zhang, Z. Y.; Pfendner, J.; Grafmüller, A.; Voth, G. A. *Biophys. J.* **2009**, *97*, 2327–2337.
- (22) Amadei, A.; Linnsen, A. B. M.; Berendsen, H. J. C. *Proteins: Struct., Funct., Genet.* **1993**, *17*, 412–425.
- (23) Saibil, H. R. *Nat. Struct. Biol.* **2000**, *7*, 711–714.
- (24) Joachim, F. *Three-dimensional electron microscopy of macromolecular assemblies*; Oxford University Press: New York, 2006.
- (25) Koch, M. H. J.; Vachette, P.; Svergun, D. I. *Q. Rev. Biophys.* **2003**, *36*, 147–227.
- (26) Lipfert, J.; Doniach, S. *Annu. Rev. Biophys. Biomol. Struct.* **2007**, *36*, 307–327.
- (27) Putnam, C. D.; Hammel, M.; Hura, G. L.; Tainer, J. A. *Q. Rev. Biophys.* **2007**, *40*, 191–285.
- (28) Linde, Y.; Buzo, A.; Gray, R. M. *IEEE Trans. Commun.* **1980**, *28*, 84–95.
- (29) Martinetz, T.; Schulten, K. *Neural Networks* **1994**, *7*, 507–522.
- (30) Wriggers, W.; Milligan, R. A.; Schulten, K.; McCammon, J. A. *J. Mol. Biol.* **1998**, *284*, 1247–1254.
- (31) Wriggers, W.; Chacón, P.; Kovacs, J.; Tama, F.; Birmanns, S. *Neurocomputing* **2004**, *56*, 165–179.
- (32) Ming, D.; Kong, Y.; Lambert, M. A.; Huang, Z.; Ma, J. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 8620–8625.
- (33) Ming, D.; Kong, Y.; Wakil, S. J.; Brink, J.; Ma, J. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 7895–7899.
- (34) Tama, F.; Wriggers, W.; Brooks, C. L. *J. Mol. Biol.* **2002**, *321*, 297–305.
- (35) Chacón, P.; Tama, F.; Wriggers, W. *J. Mol. Biol.* **2003**, *326*, 485–492.
- (36) Arkhipov, A.; Freddolino, P. L.; Imada, K.; Namba, K.; Schulten, K. *Biophys. J.* **2006**, *91*, 4589–4597.
- (37) Arkhipov, A.; Freddolino, P. L.; Schulten, K. *Structure* **2006**, *14*, 1767–1777.
- (38) Arkhipov, A.; Yin, Y.; Schulten, K. *Biophys. J.* **2008**, *95*, 2806–2821.

- (39) Yin, Y.; Arkhipov, A.; Schulten, K. *Structure* **2009**, *17*, 882–892.
- (40) Bahar, I.; Rader, A. J. *Curr. Opin. Struct. Biol.* **2005**, *15*, 586–592.
- (41) Ma, J. P. *Structure* **2005**, *13*, 373–380.
- (42) Yang, L.; Song, G.; Jernigan, R. L. *Biophys. J.* **2007**, *93*, 920–929.
- (43) Tama, F.; Valle, M.; Frank, J.; Brooks, C. L. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 9319–9323.
- (44) Wang, Y. M.; Rader, A. J.; Bahar, I.; Jernigan, R. L. *J. Struct. Biol.* **2004**, *147*, 302–314.
- (45) Trylska, J.; Tozzini, V.; McCammon, J. A. *Biophys. J.* **2005**, *89*, 1455–1463.
- (46) Kurkcuoglu, O.; Doruker, P.; Sen, T. Z.; Kloczkowski, A.; Jernigan, R. L. *Phys. Biol.* **2008**, *5*, 046005(14).
- (47) Tirion, M. M. *Phys. Rev. Lett.* **1996**, *77*, 1905–1908.
- (48) Hinsén, K.; Petrescu, A. J.; Dellerue, S.; Bellissent-Funel, M. C.; Kneller, G. R. *Chem. Phys.* **2000**, *261*, 25–37.
- (49) Atilgan, A. R.; Durell, S. R.; Jernigan, R. L.; Demirel, M. C.; Keskin, O.; Bahar, I. *Biophys. J.* **2001**, *80*, 505–515.
- (50) Lu, M. Y.; Poon, B.; Ma, J. P. *J. Chem. Theory Comput.* **2006**, *2*, 464–471.
- (51) Moritsugu, K.; Smith, J. C. *Biophys. J.* **2007**, *93*, 3460–3469.
- (52) Hinsén, K. *Bioinformatics* **2008**, *24*, 521–528.
- (53) Lyman, E.; Pfaendtner, J.; Voth, G. A. *Biophys. J.* **2008**, *95*, 4183–4192.
- (54) Hinsén, K. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, E128–E128.
- (55) Riccardi, D.; Cui, Q.; Phillips, G. N. *Biophys. J.* **2009**, *96*, 464–475.
- (56) Stember, J. N.; Wriggers, W. *J. Chem. Phys.* **2009**, *131*.
- (57) Yang, L.; Song, G.; Jernigan, R. L. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 12347–12352.
- (58) Kitao, A.; Go, N. *Curr. Opin. Struct. Biol.* **1999**, *9*, 164–169.
- (59) Berendsen, H. J. C.; Hayward, S. *Curr. Opin. Struct. Biol.* **2000**, *10*, 165–169.
- (60) Flory, P. J.; Gordon, M.; McCrum, N. G. *Proc. R. Soc. London, Ser. A* **1976**, *351*, 351–380.
- (61) Wriggers, W.; Milligan, R. A.; McCammon, J. A. *J. Struct. Biol.* **1999**, *125*, 185–195.
- (62) Wriggers, W. *Biophys. Rev.* **2010**, *2*, 21–27.
- (63) Heyd, J.; Birmanns, S. *Microsc. Today* **2008**, *16*, 6–8.
- (64) Humphrey, W.; Dalke, A.; Schulten, K. *J. Mol. Graphics* **1996**, *14*, 33–38.
- (65) Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. *J. Chem. Phys.* **1953**, *21*, 1087–1092.
- (66) Kirkpatrick, S. C. D.; Gelatt, J.; Vecchi, M. P. *Science* **1983**, *220*, 671–680.
- (67) Graceffa, P.; Dominguez, R. *J. Biol. Chem.* **2003**, *278*, 34172–34180.
- (68) Kabsch, W.; Mannherz, H. G.; Suck, D.; Pai, E. F.; Holmes, K. C. *Nature* **1990**, *347*, 37–44.
- (69) Khaitlina, S. Y.; Moraczewska, J.; Strzeleckagolaszewska, H. *Eur. J. Biochem.* **1993**, *218*, 911–920.
- (70) Pfaendtner, J.; Branduardi, D.; Parrinello, M.; Pollard, T. D.; Voth, G. A. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 12723–12728.
- (71) Chu, J. W.; Voth, G. A. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 13111–13116.
- (72) Chu, J. W.; Voth, G. A. *Biophys. J.* **2006**, *90*, 1572–1582.
- (73) Pfaendtner, J.; Lyman, E.; Pollard, T. D.; Voth, G. A. *J. Mol. Biol.* **2010**, *396*, 252–263.
- (74) Tirion, M. M.; Benavraham, D. *J. Mol. Biol.* **1993**, *230*, 186–195.
- (75) Frank, J.; Agrawal, R. K. *Nature* **2000**, *406*, 318–322.
- (76) Gabashvili, I. S.; Agrawal, R. K.; Spahn, C. M. T.; Grassucci, R. A.; Svergun, D. I.; Frank, J.; Penczek, P. *Cell* **2000**, *100*, 537–549.
- (77) Wimberly, B. T.; Brodersen, D. E.; Clemons, W. M.; Morgan-Warren, R. J.; Carter, A. P.; Vonnrhein, C.; Hartsch, T.; Ramakrishnan, V. *Nature* **2000**, *407*, 327–339.
- (78) Yusupov, M. M.; Yusupova, G. Z.; Baucom, A.; Lieberman, K.; Earnest, T. N.; Cate, J. H. D.; Noller, H. F. *Science* **2001**, *292*, 883–896.
- (79) Schuwirth, B. S.; Borovinskaya, M. A.; Hau, C. W.; Zhang, W.; Vila-Sanjurjo, A.; Holton, J. M.; Cate, J. H. D. *Science* **2005**, *310*, 827–834.
- (80) Korostelev, A.; Trakhanov, S.; Laurberg, M.; Noller, H. F. *Cell* **2006**, *126*, 1065–1077.
- (81) Selmer, M.; Dunham, C. M.; Murphy, F. V.; Weixlbaumer, A.; Petry, S.; Kelley, A. C.; Weir, J. R.; Ramakrishnan, V. *Science* **2006**, *313*, 1935–1942.
- (82) Korostelev, A.; Noller, H. F. *Trends Biochem. Sci.* **2007**, *32*, 434–441.
- (83) Sanbonmatsu, K. Y.; Joseph, S.; Tung, C. S. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 15854–15859.
- (84) Sanbonmatsu, K. Y.; Tung, C. S. *J. Struct. Biol.* **2007**, *157*, 470–480.
- (85) Gumbart, J.; Trabuco, L. G.; Schreiner, E.; Villa, E.; Schulten, K. *Structure* **2009**, *17*, 1453–1464.
- (86) Villa, E.; Sengupta, J.; Trabuco, L. G.; LeBarron, J.; Baxter, W. T.; Shaikh, T. R.; Grassucci, R. A.; Nissen, P.; Ehrenberg, M.; Schulten, K.; Frank, J. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 1063–1068.
- (87) Ermolenko, D. N.; Majumdar, Z. K.; Hickerson, R. P.; Spiegel, P. C.; Clegg, R. M.; Noller, H. F. *J. Mol. Biol.* **2007**, *370*, 530–540.
- (88) Downing, K. H.; Nogales, E. *Eur. Biophys. J. Biophys.* **1998**, *27*, 431–436.
- (89) Nogales, E.; Wolf, S. G.; Downing, K. H. *Nature* **1998**, *391*, 199–203.
- (90) Downing, K. H.; Nogales, E. *Cell Struct. Funct.* **1999**, *24*, 269–275.
- (91) Downing, K. H. *Annu. Rev. Cell. Dev. Biol.* **2000**, *16*, 89–111.
- (92) Lowe, J.; Li, H.; Downing, K. H.; Nogales, E. *J. Mol. Biol.* **2001**, *313*, 1045–1057.
- (93) Downing, K. H. *Scanning* **2003**, *25*, 74–75.
- (94) Nogales, E.; Whittaker, M.; Milligan, R. A.; Downing, K. H. *Cell* **1999**, *96*, 79–88.
- (95) Sosa, H.; Milligan, R. A. *J. Mol. Biol.* **1996**, *260*, 743–755.

- (96) Sosa, H.; Dias, D. P.; Hoenger, A.; Whittaker, M.; WilsonKubalek, E.; Sablin, E.; Fletterick, R. J.; Vale, R. D.; Milligan, R. A. *Cell* **1997**, *90*, 217–224.
- (97) Meurer-Grob, P.; Kasparian, J.; Wade, R. H. *Biochemistry* **2001**, *40*, 8000–8008.
- (98) Li, H. L.; DeRosier, D. J.; Nicholson, W. V.; Nogales, E.; Downing, K. H. *Structure* **2002**, *10*, 1317–1328.
- (99) Nogales, E.; Wolf, S. G.; Khan, I. A.; Luduena, R. F.; Downing, K. H. *Nature* **1995**, *375*, 424–427.
- (100) Keskin, O.; Durell, S. R.; Bahar, I.; Jernigan, R. L.; Covell, D. G. *Biophys. J.* **2002**, *83*, 663–680.
- (101) Gebremichael, Y.; Chu, J. W.; Voth, G. A. *Biophys. J.* **2008**, *95*, 2487–2499.
- (102) Sept, D.; Baker, N. A.; McCammon, J. A. *Protein Sci.* **2003**, *12*, 2257–2261.
- (103) Mitchison, T.; Kirschner, M. *Nature* **1984**, *312*, 237–242.
- (104) Janosi, I. M.; Chretien, D.; Flyvbjerg, H. *Eur. Biophys. J. Biophys.* **1998**, *27*, 501–513.
- (105) Lu, M. Y.; Ma, J. P. *Biophys. J.* **2005**, *89*, 2395–2401.
- (106) Tama, F.; Brooks, C. L. *Annu. Rev. Biophys. Biomol. Struct.* **2006**, *35*, 115–133.

CT100374A