

Accelerated Purification Using Generalized Nonpurifying Intermediate Functions for Large-Scale Self-Consistent Field Calculations

Jaehoon Kim and Yousung Jung*

Graduate School of EEWS (WCU) and KI for Nanocentury, Korea Advanced Institute of Science and Technology (KAIST), Daejeon 305-701, Korea

ABSTRACT: Purification is a widely used technique to calculate idempotent density matrices from a Hamiltonian in large-scale electronic structure calculations. However, the initial guess of a density matrix usually contains large errors, which require many iterations to remove them, using standard recursive schemes such as those derived by McWeeny or Holas. In this Letter, we propose a way to obtain a converged density matrix much more rapidly by removing the stability conditions that the functions have fixed points and vanishing derivatives at 0 and 1, assumptions usually made in most traditional purification methods. That is, by extending the recursive function space, which gives the approximated step function via the generalized nonpurifying intermediate functions, and optimizing them, we reduce the purification cost approximately by a factor of 1.5 compared to grand canonical purification algorithms for the linear alkanes, diamondoid, and a protein endothelin that has a very small band gap.

1. INTRODUCTION

In self-consistent field calculations such as Hartree–Fock (HF) or density functional theory (DFT), the overall computational cost is determined by the formation of the effective Hamiltonian matrix and calculation of the new density matrix from it. For the formation of Fock matrix, many efficient linear scaling algorithms have been developed, for example, continuous fast multipole methods (CFMM) and Fourier transform Coulomb (FTC) methods for Coulomb interactions^{1–6} and LinK, ONX, and multipole-accelerated methods for the HF exchange matrices.^{7–12} For an effective Hamiltonian in DFT, efficient linear scaling numerical quadrature methods for the exchange–correlation potential have also been developed.^{13,14}

The other important step in self-consistent field (SCF) calculations, updating the density matrix using the previous Fock matrix just constructed, is then usually done by diagonalizing the Fock matrix with $O(N^3)$ cost. For small- to medium-sized systems, diagonalization is not a major bottleneck due to a small prefactor of the cubic scaling. For this reason, diagonalization is a preferred choice of density update in almost all quantum chemistry calculations employing localized Gaussian basis today. For large systems, however, the cubic scaling of diagonalization becomes important and, eventually, completely dominates the overall computational cost due to a higher scaling compared to the Fock build.

Linear scaling diagonalization-free algorithms have been developed. Since there are only $O(N)$ significant matrix elements in \mathbf{P} for large enough systems, direct minimization of the energy functional with respect to the density matrix elements leads to a linear scaling algorithm.¹⁵ Alternatively, one can use the fact that the density matrix (\mathbf{P}) commutes with the Hamiltonian (\mathbf{H}) and so can be a direct function of \mathbf{H} (e.g., polynomials). In this, one essentially expresses the density matrix as a Heaviside step function of the Hamiltonian matrix:

$$\mathbf{P} = h(\mu\mathbf{I} - \mathbf{H}) \quad (1)$$

where μ is the chemical potential, \mathbf{I} is the identity matrix, and $h(\cdot)$ is Heaviside step function. One can then approximate this Heaviside step function as Fermi–Dirac¹⁶ or complementary error functions¹⁷ that can be evaluated efficiently with the Chebyshev polynomials of the Hamiltonian matrix of finite order.¹⁸ This expansion and computation only requires matrix–matrix multiplications (matmuls) and additions which can be performed with linear scaling efforts due to the locality and sparsity of the matrices involved. In addition to the direct computation of eq 1 using matrix polynomials of very high order, one can also use some recurrence formula, such as that of McWeeny¹⁹ as in eq 2, to purify an appropriately chosen initial density matrix to satisfy idempotency and ensure occupation numbers of 0's or 1's to approximate eq 1:

$$\mathbf{P}_{n+1} = 3\mathbf{P}_n^2 - 2\mathbf{P}_n^3 \quad (2)$$

We determine below that the different diagonalization-free algorithms described above (minimization, expansion, or purification) have different convergence behaviors. For the Fermi operator expansion (FOE) method, error decreases as $O(\epsilon^m)$, where m is the number of required matmuls related to the degree of polynomial, while for an improved FOE algorithm with a fast summation technique, the error decreases more quickly as $O(\epsilon^{m^2/4})$. We find a similar convergence rate to the FOE method, i.e., $O(\epsilon^{cm})$ for some c , derived for the Curvy-step (CS) method²⁰ since it uses Baker–Hausdorff expansion, which is also a type of matrix polynomial expansion, but with a reduced cost due to a better initial guess. For purification algorithms, if we write the density matrix at the n th step as $\mathbf{P}_n = \mathbf{P} + \epsilon_n$ where \mathbf{P} is the exact density matrix, then $\mathbf{P}_{n+1} = \mathbf{P} + (3\mathbf{I} - 6\mathbf{P})\epsilon_n^2 + O(\epsilon_n^3)$ with \mathbf{P} and ϵ_n commuting, where $O(\epsilon_n^3)$ indicates a matrix whose elements are on the same order of magnitude as those of ϵ_n^3 (see appendix B in ref 21). With the initial guess of density matrix specially chosen as eq 3, the absolute value of eigenvalues of ϵ_n is always lower than 0.5, and thus $\mathbf{P}_\infty = \mathbf{P}$. This initial guess \mathbf{P}_0 of eq 3 is

Published: October 28, 2011

specially designed and scaled for the eigenvalues to lie between the $[0,1]$ interval in reverse order as part of the purification procedure to calculate the matrix step function. Therefore, the $\text{Tr}(\mathbf{P}_0)$ does not mean that it is the true number of electrons in the system, but rather will be converged to the correct number of electrons during the purification iterations.

$$\mathbf{P}_0 = \frac{\lambda}{2}(\mu\mathbf{I} - \mathbf{H}) + \frac{1}{2}\mathbf{I} \quad (3)$$

$$\lambda = \min\left(\frac{1}{H_{\max} - \mu}, \frac{1}{\mu - H_{\min}}\right) \quad (4)$$

where H_{\max} and H_{\min} are the upper and lower bounds of the eigenvalue spectrum of the Hamiltonian.²¹ The errors in the purification algorithm of McWeeny, for example, decrease as $\mathbf{P} - \mathbf{P}_n = O(\epsilon^{2^n}) = O(\epsilon^{2^{m/2}})$.

The computational cost to achieve a given accuracy can then be written as

$$\text{cost} = O(f^{-1}(\text{thresh})) \quad (5)$$

where the error is $O(\epsilon^{f(m)})$ and thresh defines the accuracy with $\text{maxerror} = 10^{-\text{thresh}}$. For the curvy-step and FOE methods, the cost is linearly dependent on thresh. For an improved FOE method with the fast summation technique, $\text{cost} = O((\text{thresh})^{1/2})$, and for purification methods, $\text{cost} = O(\log(\text{thresh}))$. Consideration of these different scaling properties of error reduction for various density matrix computation methods indicates that purification is a computationally more efficient way to calculate the density matrix from the Hamiltonian to achieve a given accuracy compared to other indirect diagonalization approaches. A similar and more detailed comparison of various density matrix calculation methods has also been given recently, leading to a similar conclusion that the recursive polynomial expansion is generally more efficient than other minimization methods.²² This Letter is about reducing the computational cost of existing purification algorithms even further by lifting one of the usual assumptions traditionally made.

In purification, one can of course go beyond third-order in eq 2 for faster convergence (fewer iterations) toward the true density matrix. When using the higher-order functions, however, one must consider the cost to compute the polynomial itself at each iteration. Recently, we have shown²³ that the optimal purification scheme is obtained when the fifth-order polynomial of Holas²⁴ is repeatedly applied throughout all iterations, and either the use of McWeeny or higher-order polynomials throughout or the mixed use of different-order polynomials at different iterations results in higher cost.

Regardless of the forms and degrees of polynomials employed in the purification algorithm, however, nearly all traditional purification methods generally assume a convention that at each iteration the polynomials are required to purify the density matrix; that is, the polynomials have stable fixed points at 0 and/or 1. However, this procedure of using the true purification functions at all iterations is not a necessary condition since only the final converged density matrix needs to be idempotent. In other words, if some generalized polynomials, even if they are nonpurifying functions, can reduce the large errors in density matrix more rapidly than the traditional true purification polynomials, one can consider using these nonpurifying functions during early and intermediate stages of iterations and apply the true purification functions, such as those of McWeeny or Holas, only at the last stages to ensure idempotency.

Table 1. Optimized Parameters for Nonpurifying Intermediate Functions Defined in eq 6 for Each Iteration in OP_s^a

	iteration 1	iteration 2	iteration 3	iteration 4
a_1	16.05230180	18.01787054	10.80499078	7.284154647
a_2	0.486483492	0.394957751	0.417919928	0.315415352
c	0.3515			

^a c in the last row determines the usable boundary of input eigenspectrum defined in eqs 7–9.

We call, in this paper, these generalized nonpurifying functions specially designed and optimized to accelerate the convergence “intermediate functions” (MFs), to be distinguished from a true purification function (PF). To this end, we here suggest an optimized purification (OP) algorithm that can reduce the total purification cost in terms of matmuls approximately by a factor of 1.5 by combining the generalized nonpurifying MFs and purifying PFs, as compared to the conventional schemes that usually use PFs repeatedly until the convergence is reached, such as canonical purification (CP) or grand canonical purification (GCP).

2. THEORY

The key aspect of our approach is that we remove the stability condition that the functions have stable fixed points at 0 and 1 at each iteration, and we adjust functions differently for each iteration. That is, we extend the recursive function space, which gives the approximated step function via the parametrized functions. We restrict the function space to be explored using three constraints:

1. maintaining symmetry with respect to the point (0.5,0.5)
2. using the fifth-order polynomials throughout all iterations
3. passing through the two fixed points (1,1) and (0,0).

We use constraint 1 to conserve symmetry during iterations since the shifted step function has a symmetry around (0.5,0.5) and the number of matmuls required to evaluate the functions can be reduced due to symmetry. The reduction in cost due to symmetry arises since the polynomial with symmetry can be expanded with either an odd or even function.²⁴ Constraint 2 is imposed to minimize the computational cost of polynomial evaluation, motivated by our recent analytical proof²³ that the repeated application of the fifth-order Holas polynomial throughout all iterations is an optimal scheme for the density matrix purification when polynomials have stable fixed points at 0 and 1. Here, we extend this theorem to the nonpurifying polynomials. Constraint 3 is needed to ensure that, once purified to 0's and 1's, they remain as 0's and 1's even with the application of nonpurifying functions. Since all of the MFs must be applied in sequence as one “quanta” in OP, without constraint 3, application of many nonpurifying functions may yield the matrix that diverges.

The functional form that satisfies all three conditions described above is

$$f(x, a_1, a_2) = x + a_1x(x-1)\left(x - \frac{1}{2}\right)(x^2 - x - a_2) \quad (6)$$

Since we have only two adjustable parameters, a_1 and a_2 , we search for the optimized functions at each iteration directly, as described below. If the fifth-order Holas purification process

1. Input : Hamiltonian (\mathbf{H})
2. Upper and lower bounds of the eigenspectrum of Hamiltonian
3. Calculate Initial guess,

$$\mathbf{P}_0 = \frac{\lambda}{2}(\mu\mathbf{I} - \mathbf{H}) + \frac{1}{2}\mathbf{I} \text{ with } \lambda = \min\left(\frac{1.7030}{H_{max}-\mu}, \frac{1.7030}{\mu-H_{min}}\right)$$

4. Compute the intermediate functions in sequence, i.e.,

$$\mathbf{P}_4 = F_4\left(F_3\left(F_2\left(F_1\left(\mathbf{P}_0\right)\right)\right)\right).$$

5. If (convergence I)

If (convergence II)

Stop (Purification completed)

Else

Purify density matrix using the true purification function.

$$\mathbf{P}_{k+1} = \mathbf{P}_k^3 [10 - 15\mathbf{P}_k + 6\mathbf{P}_k^2], k \geq 4$$

6. Else go to step 4 (\mathbf{P}_0 is replaced by \mathbf{P}_4)

$$\text{Convergence criterion I : } c_1 = \frac{\|\mathbf{P}_4 - \mathbf{P}_0\|_F}{\sqrt{N}} = \sqrt{\frac{\text{Tr}((\mathbf{P}_4 - \mathbf{P}_0)^2)}{N}} < 10^{-2}$$

$$\text{Convergence criterion II : } c_2 = \max[(\mathbf{P}_{k+1} - \mathbf{P}_k)_{ij}] < 10^{-9}$$

Figure 1. Flowchart of the optimized purification (OP) algorithm. $\|\cdot\|_F$ is a Frobenius norm defined as $\|\mathbf{A}\|_F = (\text{Tr}(\mathbf{A}^T\mathbf{A}))^{1/2}$, where N is a dimension of the density matrix and $\text{Tr}()$ denotes the trace of a matrix.

were indeed optimal as in ref 23, one should recover $(a_1, a_2) = (6, 1/3)$ for all iterations in this extended function space when the parameters are optimized to produce the best approximate step function. The fact that the optimal values of (a_1, a_2) in eq 6 turn out to be different from those of Holas (as shown in Table 1) demonstrates that Holas' polynomial is not optimal if we generalize the function space to nonpurifying functions during iterations. A substantial improvement in convergence rate is indeed obtained with this generalization, as shown below. We abbreviate the proposed optimized purification scheme using the fifth-order polynomials as OP_5 .

3. OPTIMIZATION

The number of intermediate purification functions (MFs), M , is chosen considering three factors: (i) The more efficient sequence of polynomials may be obtained with the larger number of MFs that are optimized; however, (ii) since all of the optimized MFs must be applied in order as one packet or one "quanta" (see algorithmic flowchart), the smaller number of M is more flexible to apply, and (iii) optimization of the parameters themselves becomes exponentially more difficult as M increases. On the basis of this trade-off between the optimized performance and flexibility, and also since a purification process for calculating a density matrix typically requires at least four iterations, we set the number of optimized MFs to be $M = 4$. In principle, one can go beyond this, if necessary. We comment on this aspect at the end of the present Letter.

The region of adjustable parameter space explored is set to be inside a rectangular space whose edges are $(-6, 0)$ and $(18, 2/3)$. This parameter space has a center at $(6, 1/3)$ which recovers Holas' fifth-order formula. To find the global minimum for (a_1, a_2) using the error function defined below, we explored the entire parameter space using a dual-resolution, cruder spacing $(1.5, 1/9)$ and a finer spacing $(10^{-10}, 10^{-10})$. All $(17 \times 7)^4$ possibilities for

$M = 4$ were considered to find the local minima at the lower resolution, which were then subsequently searched for a global minimum with a finer grid.

We define the error function to determine the optimal parameters as follows:

$$e[F(x), \alpha] = \frac{\int_{0.5}^{1+c} |F(x) - h(x - 0.5)|^\alpha dx}{0.5 + c} \quad (7)$$

where $F(x)$ is a generalized nonpurifying intermediate function (MF) that reduces the errors in the density matrix rapidly, i.e.,

$$F_0(x) = x \quad (8a)$$

$$F_i(x) = f(F_{i-1}(x), a_1(i), a_2(i)) \text{ for } i = 1 \sim 4 \quad (8b)$$

$$F(x) = F_4(x) \quad (8c)$$

The α can be any positive real number, and c is a parameter that extends the purification region beyond $[0, 1]$ and is determined by the minimal positive number that satisfies

$$\min_c : [F(1 + c) - 1]^2 > z \quad (9)$$

In eq 9, z is a tolerance parameter that sets a small allowed variation of occupancy around the exact occupancy, 0 or 1. We used $z = 0.003$. For α , $\alpha = 2$ may be the most natural choice in a least-squares sense; however, we regarded this exponent of the error function also as a variable for a finer optimization. By generating uniformly distributed random numbers for α in the interval $[1, 3]$ and finding an α that minimizes the error function in eq 7, we obtained and used $\alpha = 1.7199$.

The desirable properties of the optimized MFs would include (i) a large slope near the chemical potential, (ii) a wideness of eigenvalue spectrum (input) that can be covered by the chosen intermediate functions, and (iii) a narrowness of the eigenvalue spectrum (output) of the processed density matrix. The proposed error function, eq 7, is designed to meet these three conditions as much as possible. First, for i, the larger slope at the chemical potential would better approximate the step function and, hence, would require a smaller number of iterations (in particular for smaller band gap problems). Second, for ii, since $F(\mathbf{P})$ is a polynomial of \mathbf{P} , its approximation of the step function is only applicable for a limited region of the eigenvalue spectrum, and the parameter c in eq 9 determines this boundary. In our approach, the approximate step function has an extended usable interval $[-c, 1 + c]$ instead of $[0, 1]$.

4. RESULTS

The optimized nonpurifying MFs are summarized in Table 1, and the proposed accelerated purification algorithm is summarized in Figure 1 as a flowchart. Purification begins by applying the optimized nonpurifying MFs to the appropriately chosen initial density matrix (four iterations in one sweep) followed by the first convergence check. If the errors are still large, e.g., if the Frobenius norm of the difference between the density matrix and its square is more than $N \times (10^{-2})^2$, where N is a dimension of the density matrix, one goes back and keeps applying multiple sweeps of MFs until the criterion of convergence I is met. Since the functions are designed to make the spectral norm of errors for the density matrix near 10^{-2} , the repeated iteration of our intermediate functions must meet convergence I. Once

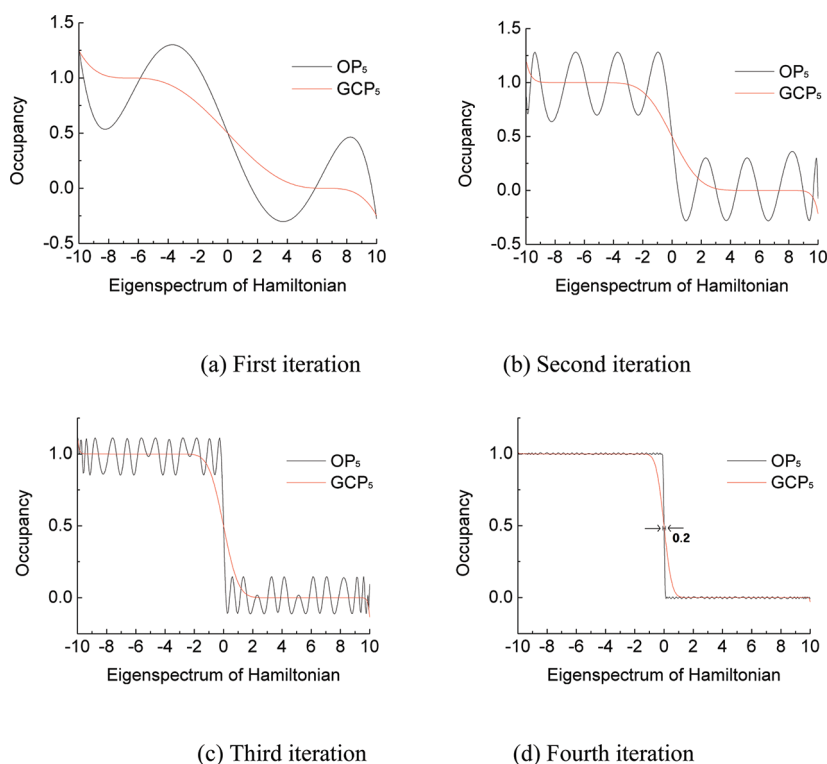


Figure 2. Efficiency of the OP scheme (OP_5) proposed here using the sequence of optimized fifth-order polynomials, compared to that of GCP using Holas' fifth-order function (GCP_5), along the first four iterations. The x -axis is the eigenspectrum of the Hamiltonian for a toy system with the spectral width $\Delta\varepsilon = 20$ and the chemical potential $\mu = 0$.

convergence criterion I is met, one then switches to a standard true purification function, i.e., the fifth-order Holas polynomial, and keeps applying the same Holas function until convergence criterion II ($<10^{-9}$) is satisfied.

The numerical results for the evenly distributed eigenvalue spectrum of the Hamiltonian matrix in $[-10,10]$ are shown in Figure 2, where we compare OP_5 and GCP_5 , where the subscript 5 means that the fifth-order nonpurifying or Holas polynomials are used, respectively. At first, OP_5 appears to have larger errors than GCP_5 during the initial 1–3 iterations, but these nonpurifying MFs in OP_5 play an important role in the rapid reduction of large errors so that the subsequent intermediate function stabilizes small errors more uniformly and effectively. As a result, one can visually see a clear improvement of OP_5 over GCP_5 after four iterations (1 sweep of OP_5). The net effect of applying these generalized optimized functions (that are not intended to purify the density matrix) is a small wiggling of the occupation numbers near 0 and 1, which needs to be purified subsequently. Usually, 1–3 applications of the true purification function are found to be sufficient to reduce the maximum error below the target criterion II.

The proposed OP method has a few advantages over the existing purification schemes using the true purification functions. First, the valid purification region for the fifth-order Holas polynomial is formally $[-0.259, 1.259]$, although traditionally only the limited interval $[0,1]$ was utilized in density matrix calculation. In OP, the applicable range of occupation numbers is wider, $[-0.3515, 1.3515]$. Second, in a graph similar to Figure 2, the slope or the width of a “rising region” at the chemical potential can serve as an indicator of accuracy for the approximate step function. Here, we use the latter, i.e., a minimal range

where the occupancy is between 10^{-2} and $1 - 10^{-2}$. Under this measure, we reach the width 0.5 ± 0.009 after four iterations in the case of OP, which is effectively 0.5 ± 0.005 when the initial guess is scaled properly. Holas' fifth-order purification, however, yields the effective width of the rising region to be 0.5 ± 0.052 (about a factor of 10 larger than OP) after four iterations and takes eight iterations total to obtain the same accuracy as OP (width being ± 0.005).

We implemented the proposed OP algorithm into the development version of Q-CHEM²⁵ and assessed its practical performance for chemical systems, namely, linear alkane and diamond structures, using the BLYP density functional calculations at the STO-3G and 6-31G** bases. The symmetric transformation, i.e., $S^{-1/2}$, where S is overlap matrix, was used for orthogonalization of the atomic orbital (AO) basis. In Tables 2 and 3, we compare the performance of OP and other expansion and purification methods in the literature in terms of the average number of matmuls required per SCF cycle. For the present calculations, we used the superposition of atomic density for an initial guess and Pulay's DIIS algorithm with five subspaces for the SCF algorithm. As shown in Tables 2 and 3, the OP algorithm is faster than the existing expansion or purification methods by roughly a factor of 1.5.

To test the performance of OP for smaller band gap systems, we additionally considered a protein called endothelin that has 328 atoms (or 171 non-hydrogen atoms). For comparison, $C_{240}H_{482}$ (alkane) and $C_{100}H_{100}$ (diamondoid) have HOMO–LUMO (highest occupied molecular orbital–lowest unoccupied molecular orbital) gaps of 0.877 and 0.310 hartree, respectively, at the BLYP/STO-3G level, while endothelin has a HOMO–LUMO gap of 0.039 hartree at the same level.

Table 2. Average Number of Matmuls Needed to Compute the Density Matrix Per SCF Cycle for Alkane Chains of Varying Size^a

method	STO-3G				6-31G**
	C ₆₀ H ₁₂₂	C ₁₂₀ H ₂₄₂	C ₁₈₀ H ₃₆₂	C ₂₄₀ H ₄₈₂	C ₆₀ H ₁₂₂
CP (ref 20)	33 (6)	33 (5)	33 (5)	33 (5)	64 (7)
CS (ref 20)	34 (7)	34 (7)	35 (7)	35 (7)	49 (9)
FOE (ref 17)	27	27	27	27	47
GCP ₃	30 (6)	30 (5)	30 (5)	30 (5)	40 (7)
GCP ₅	30 (6)	30 (5)	30 (5)	30 (5)	33 (7)
OP ₃	23 (6)	23 (5)	23 (5)	23 (5)	31 (7)
OP ₅	22 (6)	22 (5)	22 (5)	22 (5)	25 (7)

^aFor GCP and OP, the average number of matmuls per SCF cycle is calculated to be the same as the number of matmuls for the last SCF cycle for these molecules. BLYP/STO-3G or BLYP/6-31G** calculations were performed, and the numbers in parentheses are the total number of SCF cycles taken for SCF convergence. For OP, the extra matmuls to calculate the Frobenius norm in convergence criterion I are also taken into account.

Endothelin also has a spectral width of 87.8 hartree, yielding $\xi/\Delta\epsilon = 0.0004$ at the BLYP/STO-3G level, which exceeds the $\xi/\Delta\epsilon = 0.01$ ratio for which OP is originally designed. For this small-gap protein, we find that the number of matmuls required for the purification of the last SCF cycle is 35 for OP₅, still yielding a significant speedup by a factor of 1.6 when compared with 54 matmuls for both GCP₅ and GCP₃. Therefore, it demonstrates the key advantage of OP for a fast error reduction during the early iterations of purification regardless of the band gap. Switching to a true purification function (Figure 1) such as Holas' then allows OP to handle molecular systems whose band gaps are beyond the $\xi/\Delta\epsilon = 0.01$ limit.

To numerically assess the previous mathematical proof (and current extension without proof) that the fifth-order polynomial is optimal for purifying the density matrix, we compared the performance (the number of matmuls taken to achieve a target accuracy) of the OP algorithm proposed here with degrees 3 and 5. For a fair comparison with OP₅ where the four fifth-order MFs were optimized, we used six MFs for the third-order polynomials in OP₃, noting that the application of the four fifth-order MFs and six third-order MFs requires the same number of matmuls to evaluate, namely, 12. The optimized parameters for OP₃ are summarized in Table 4. We used the same optimization scheme to find the optimal third-order polynomials as in the fifth-order functions. The results in Tables 2 and 3 show that the cost of matrix computations for OP₅ is slightly lower than that of OP₃ for selected alkane chain and diamond structures. The number of matmuls for the GCP₃ calculation (using McWeeny) is also compared with that of GCP₅ (using the fifth-order Holas polynomial), whose results show that GCP₅ is indeed marginally more efficient than GCP₃. However, the result for a diamond structure at the STO-3G basis, where GCP₅ required one more matmul than GCP₃ on average, is perhaps due to a fact that three matmuls must be used to evaluate the fifth-order Holas polynomial, while two are required in the case of the McWeeny polynomial.

In the final stage of completing this manuscript, we became aware of Rubensson's scaled purification (SP) method,²⁶ which performs impressively compared to existing purification

Table 3. Average Number of Matrix Multiplications Per SCF Cycle for H-Terminated Diamond Structure C₁₀₀H₁₀₀^a

method	STO-3G	6-31G**
GCP ₃	32 (7)	45 (9)
GCP ₅	33 (7)	43 (9)
OP ₃	23 (7)	37 (9)
OP ₅	22 (7)	35 (9)

^aBLYP/STO-3G or BLYP/6-31G** calculations were performed. For OP, the extra matmuls to calculate the Frobenius norm in convergence criterion I are also taken into account. The average number of matmuls per SCF cycle is calculated to be the same as the number of matmuls for the last SCF cycle for this molecule.

Table 4. Optimized Nonpurifying Intermediate MFs at Each Iteration in OP₃ That Consists of Six Third-Order Polynomials^a

	iteration 1	iteration 2	iteration 3	iteration 4	iteration 5	iteration 6
<i>a</i>	-0.51788	-5.23743	-7.90017	-2.83308	-6.06444	-1.51465

^aThe functional form takes $f(x,a) = a(x-1)(x-1/2)x + x$.

methods, and we comment on it here. The main idea in SP is to stretch out the eigenspectrum of the initial density matrix and fold back to the desired [0,1] interval to accelerate the error reduction. To compare SP (algorithm 1 of ref 26) and OP, proposed here, we considered a toy model with $\xi/\Delta\epsilon = 0.01$ (where $\Delta\epsilon$ is a spectral width of the Hamiltonian and ξ is a band gap). This band gap is chosen since it is the affordable band gap limit of OP. We compared the errors of OP₅ after four iterations and SP after six iterations, which require the same computational work. OP₅ has an error $\|\mathbf{P} - \mathbf{P}_4\|_2 \approx 0.008$ after four iterations, where $\|\cdot\|_2$ is the spectral norm, while the SP algorithm shows more wiggings with $\|\mathbf{P} - \mathbf{P}_6\|_2 \approx 0.008$ after six iterations. Of course, it should also be noted that, due to a greater flexibility of the SP algorithm compared to OP, which must be iterated with the multiples of *m*, i.e., 4 in the present case, the relative performance in practice must yet be seen. The practical performance comparison will also depend on the band gap of a particular system of interest with a certain spectral width. For example, for a smaller gap ratio ($\xi/\Delta\epsilon = 0.005$), the SP algorithm has a smaller error (0.091) than OP (0.165), although it is the beyond the affordable band gap ratio for which OP is designed. In principle, one can systematically design a customized set of optimized intermediate functions in OP for different band gap problems by varying the number of MFs and reoptimizing them depending on the band gap.

5. CONCLUSIONS

We have shown that the computational cost of approximating a matrix step function in density matrix calculations can be substantially reduced by optimizing the recursive polynomials used at each iteration differently in an extended function space. In particular, we proposed the use of generalized nonpurifying intermediate functions that are optimized to reduce large errors in the initial density matrix rapidly, which are then subsequently purified to exact idempotency using a usual purification formula, such as the fifth-order Holas polynomial. Numerical assessments using density functional calculations for the linear alkane,

diamond, and a protein endothelin demonstrate that the cost of purification is reduced by a factor of 1.5 in terms of the total number of matmuls by using the proposed OP scheme as compared to the CP or GCP algorithms, regardless of the band gap. The optimal number of intermediate functions, e.g., four for OP_s, however, may depend on the band gap of a problem of interest. Although our optimized purification is developed mainly for a fast computation of the matrix step function from a given Hamiltonian, since purification is a widely used technique in electronic structure theory in various contexts, exploring its advantage in contexts such as in theories that use purification to guide unconstrained density matrix minimization²⁷ may be an interesting subject of future study.

AUTHOR INFORMATION

Corresponding Author

*E-mail: ysjn@kaist.ac.kr.

ACKNOWLEDGMENT

We are pleased to acknowledge the support of Basic Science Research (2010-0023018) and WCU (R-31-2008-000-10055-0) programs funded by the Ministry of Education, Science, and Technology of Korea. We acknowledge the generous computing time from KISTI (KSC-2011-G1-01).

REFERENCES

- (1) White, C. A.; Johnson, B. G.; Gill, P. M. G.; Head-Gordon, M. The continuous fast multipole method. *Chem. Phys. Lett.* **1994**, *230*, 8–16.
- (2) Strain, M. C.; Scuseria, G. E.; Frisch, M. J. Achieving Linear Scaling for the Electronic Quantum Coulomb Problem. *Science* **1996**, *271*, 51–53.
- (3) White, C. A.; Johnson, B. G.; Gill, P. M. G.; Head-Gordon, M. Linear scaling density functional calculations via the continuous fast multipole method. *Chem. Phys. Lett.* **1996**, *253*, 268–278.
- (4) White, C. A.; Head-Gordon, M. Rotating around the quartic angular momentum barrier in fast multipole method calculations. *J. Chem. Phys.* **1996**, *105*, 5061.
- (5) Challacombe, M.; Schwegler, E. Linear scaling computation of the Fock matrix. *J. Chem. Phys.* **1997**, *106*, 5526–5536.
- (6) Shao, Y.; Head-Gordon, M. An improved J matrix engine for density functional theory calculation. *Chem. Phys. Lett.* **2000**, *323*, 425–433.
- (7) Schwegler, E.; Challacombe, M. Linear scaling computation of the Hartree-Fock exchange matrix. *J. Chem. Phys.* **1996**, *105*, 2726–2734.
- (8) Burant, J. C.; Scuseria, G. E.; Frisch, M. J. A linear scaling method for Hartree–Fock exchange calculations of large molecules. *J. Chem. Phys.* **1996**, *105*, 8969.
- (9) Schwegler, E.; Challacombe, M.; Head-Gordon, M. Linear scaling computation of the Fock matrix. II. Rigorous bounds on exchange integrals and incremental Fock build. *J. Chem. Phys.* **1997**, *106*, 9708.
- (10) Ochsenfeld, C.; White, C. A.; Head-Gordon, M. Linear and sublinear scaling formation of Hartree–Fock-type exchange matrices. *J. Chem. Phys.* **1998**, *109*, 1663.
- (11) Schwegler, E.; Challacombe, M. Linear scaling computation of the Fock matrix. IV. Multipole accelerated formation of the exchange matrix. *J. Chem. Phys.* **1999**, *111*, 6223.
- (12) Schwegler, E.; Challacombe, M. Linear scaling computation of the Fock matrix. III. Formation of the exchange matrix with permutational symmetry. *Theor. Chem. Acc.* **2000**, *104*, 344–349.
- (13) Pérez-Jorda, J. M.; Yang, W. An algorithm for 3D numerical integration that scales linearly with the size of the molecule. *Chem. Phys. Lett.* **1995**, *241*, 469–476.
- (14) Stratmann, R. E.; Scuseria, G. E.; Frisch, M. J. Achieving linear scaling in exchange-correlation density functional quadratures. *Chem. Phys. Lett.* **1996**, *257*, 213–223.
- (15) Li, X. P.; Nunes, R.; Vanderbilt, D. Density-matrix electronic-structure method with linear system-size scaling. *Phys. Rev. B* **1993**, *47*, 10891.
- (16) Goedecker, S.; Teter, M. Tight-binding electronic-structure calculations and tight-binding molecular dynamics with localized orbitals. *Phys. Rev. B* **1995**, *51*, 9455.
- (17) Liang, W.; Saravanan, C.; Shao, Y.; Baer, R.; Bell, A. T.; Head-Gordon, M. Improved Fermi operator expansion methods for fast electronic structure calculations. *J. Chem. Phys.* **2003**, *119*, 4117.
- (18) Paterson, M. S.; Stockmeyer, L. J. On the Number of Nonscalar Multiplications Necessary to Evaluate Polynomials. *SICOMP* **1973**, *2*, 60.
- (19) McWeeny, R. Some recent advances in density matrix theory. *Rev. Mod. Phys.* **1960**, *32*, 335.
- (20) Shao, Y.; Saravanan, C.; Head-Gordon, M.; White, C. A. Curvy steps for density matrix-based energy minimization: Application to large-scale self-consistent-field calculations. *J. Chem. Phys.* **2003**, *118*, 6144.
- (21) Palsler, A.; Manolopoulos, D. Canonical purification of the density matrix in electronic-structure theory. *Phys. Rev. B* **1998**, *58*, 12704–12711.
- (22) Rudberg, E.; Rubensson, E. H. Assessment of density matrix methods for linear scaling electronic structure calculations. *J. Phys.: Condens. Matter* **2011**, *23*, 075502.
- (23) Kim, J.; Jung, Y. On the optimal symmetric purification scheme of the one-particle density matrix. *Chem. Phys. Lett.* **2011**, *511*, 159–160.
- (24) Holas, A. Transforms for idempotency purification of density matrices in linear-scaling electronic-structure calculations. *Chem. Phys. Lett.* **2001**, *340*, 552–558.
- (25) Shao, Y.; Molnar, L. F.; Jung, Y.; Kussmann, J.; Ochsenfeld, C.; Brown, S. T.; Gilbert, A. T. B.; Slipchenko, L. V.; Levchenko, S. V.; O'Neill, D. P.; DiStasio, R. A.; Lochan, R. C.; Wang, T.; Beran, G. J. O.; Besley, N. A.; Herbert, J. M.; Lin, C. Y.; Van Voorhis, T.; Chien, S. H.; Sodt, A.; Steele, R. P.; Rassolov, V. A.; Maslen, P. E.; Korambath, P. P.; Adamson, R. D.; Austin, B.; Baker, J.; Byrd, E. F. C.; Dachselt, H.; Doerksen, R. J.; Dreuw, A.; Dunietz, B. D.; Dutoi, A. D.; Furlani, T. R.; Gwaltney, S. R.; Heyden, A.; Hirata, S.; Hsu, C.; Kedziora, G.; Khalliulin, R. Z.; Klunzinger, P.; Lee, A. M.; Lee, M. S.; Liang, W.; Lotan, L.; Nair, N.; Peters, B.; Proynov, E. I.; Pieniazek, P. A.; Rhee, Y. M.; Ritchie, J.; Rosta, E.; Sherrill, C. D.; Simmonett, A. C.; Subotnik, J. E.; Woodcock, H. L.; Zhang, W.; Bell, A. T.; Chakraborty, A. K.; Chipman, D. M.; Keil, F. J.; Warshel, A.; Hehre, W. J.; Schaefer, H. F.; Kong, J.; Krylov, A. I.; Gill, P. M. W.; Head-Gordon, M. Advances in methods and algorithms in a modern quantum chemistry program package. *Phys. Chem. Chem. Phys.* **2006**, *8*, 3172–91.
- (26) Rubensson, E. H. Non-monotonic recursive polynomial expansions for linear scaling calculation of the density matrix. *J. Chem. Theory Comput.* **2011**, *7*, 1233–1236.
- (27) Millam, J. M.; Scuseria, G. E. Linear scaling conjugate gradient density matrix search as an alternative to diagonalization for first principles electronic structure calculations. *J. Chem. Phys.* **1997**, *106*, 5569.

Improving MM-GB/SA Scoring through the Application of the Variable Dielectric Model

Krishna Ravindranathan,[†] Julian Tirado-Rives,[†] William L. Jorgensen,[†] and Cristiano R. W. Guimarães^{*,†}

[†]Department of Chemistry, Yale University, New Haven, Connecticut 06520, United States

^{*}Worldwide Medicinal Chemistry Department, Pfizer Global Research and Development, Eastern Point Road, Groton, Connecticut 06340, United States

 Supporting Information

ABSTRACT: A variable dielectric model based on residue types for better description of protein–ligand electrostatics in MM-GBSA scoring is reported. The variable dielectric approach provides better correlation with binding data and reduces the score dynamic range, typically observed in the standard MM-GB/SA method. The latter supports the view that exaggerated enthalpic separation between weak and potent compounds due to the lack of shielding effects in the model is greatly responsible for the wide scoring spread.

INTRODUCTION

Molecular mechanics based scoring methods using all atom force fields coupled with Poisson–Boltzmann (MM-PB/SA)¹ or generalized Born calculations (MM-GB/SA)² to model solvation have recently seen an upsurge in popularity. When compared to docking scoring functions, the physics-based methods provide improved enrichment in the virtual screening of databases and better correlation between calculated binding affinities and experimental data.³ Their main purpose is to rescore docking poses in order to circumvent limitations of the docking scoring functions, in particular the ones associated with desolvation, intramolecular, and entropy penalties for the ligands upon binding.

We investigated the performance of our own flavor of MM-GB/SA when rescoring docking poses of congeneric series for pharmaceutically relevant targets.⁴ The correlations with experimental results obtained with the physics-based scoring were far superior to the ones obtained with the Glide XP scoring function and competitive with the computationally intensive FEP methods.^{4,5} Despite showing good accuracy when applied within a series, much work is necessary to improve the MM-GB/SA method and gain greater efficiency in drug design. For example, MM-GB/SA suffers from poor estimation of protein desolvation and a large dynamic range observed in the scoring when compared to the experimental range.

In the case of the GB/SA protein desolvation, substituting this term with the free energy associated with displacing binding-site waters upon ligand binding estimated by the WaterMap method,⁶ which treats the solvent explicitly, provides superior results.^{4b,7} As for the large theoretical dynamic range for the binding energies, that seems to be a direct result of the degree of sampling since FEP simulations with restricted flexibility as well as MM-GB/SA approaches that make use of a single configuration for the protein–ligand complexes are plagued by this.^{4a} Computational van't Hoff analysis suggests that the wider scoring spread is affected by not only missing entropic contributions due to restricted

sampling but also exaggerated enthalpic separation between compounds.^{4a}

One plausible cause for the exaggerated enthalpic gap is the application of an internal dielectric constant (ϵ_{in}) of 1 in a model where protein motions and polarization are not taken into account. Hence, electrostatic interactions are not shielded enough, and protein–ligand electrostatic attractions and repulsions are overestimated, causing the large separation between potent and weak compounds.^{4a,b} As previously described, when the protein permanent dipoles are included explicitly but their relaxation, i.e., the protein reorganization, and the protein induced dipoles are considered implicitly, the value of ϵ_{in} is not well-defined.⁸ Warshel and co-workers suggest that ϵ_{in} should be between 4 and 6 for dipole–charge interactions and 10 for charge–charge interactions.^{9,10} More recently, a variable dielectric model has been developed to increase the accuracy in protein side chain and loop predictions.¹¹ The authors introduced an energy model where ϵ_{in} is allowed to vary as a function of the interacting residues.

In this work, we explore the use of a variable dielectric model based on residue types to alleviate the overestimation of electrostatic effects between protein residues and ligands for improved MM-GBSA scoring. Since a poor description of protein–ligand electrostatic interactions could not only result in a wider scoring spread but also affect the correlation with experimental results, we decided to use binding data to derive the set of variable dielectric constants. Specifically, the pharmaceutically relevant targets CDK2, FactorXa, p38, and PDE10A and respective congeneric series were considered in the optimization process that led to the set of variable dielectric constants, subsequently tested on two additional data sets, the human carbonic anhydrase (hCAII) and a second p38 chemical series.

Received: August 12, 2011

Published: November 14, 2011

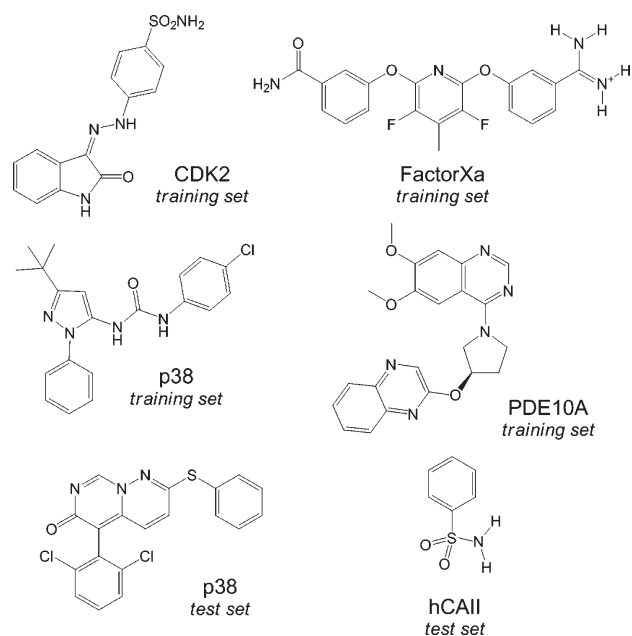


Figure 1. Representative structures for the congeneric series used in this work. Tables with all derivatives and binding data can be found in the Supporting Information.

METHODS

Data Sets and System Setup. The crystal structures for CDK2 (PDB ID: 1E9H), FactorXa (PDB ID: 1FJS), PDE10A (unpublished structure), both p38 chemical series (PDB IDs: 1KV1 and 1OUY), and hCAII (PDB ID: 2WEJ) were selected. The 1E9H CDK2 structure was modified as described before,^{4c} as it is not consistent with the conditions for the biological assay. All complexes were submitted to a series of restrained, partial minimizations using the OPLS_2005 force field¹² within MacroModel.¹³ Prior to the docking calculations using Glide XP,¹⁴ the selected congeneric series of CDK2,¹⁵ FactorXa,¹⁶ PDE10A,¹⁷ p38,^{18,19} and hCAII²⁰ were submitted to a pre-energy minimization using the OPLS_2005 force field and the GB/SA method as the implicit water model.²¹ Representative structures for each chemical series are illustrated in Figure 1. Tables with all derivatives and binding data can be found in the Supporting Information. In order to accommodate for the fact that the protein structure used for docking will not in general be optimized to fit a particular ligand, the van der Waals radii for nonpolar protein atoms were scaled by a factor of 0.8, while those for the ligands were not scaled.

MM-GB/SA Rescoring. Our implementation of the MM-GB/SA rescoring procedure (eq 1) has been described in detail elsewhere.⁴ In eq 1, ΔE_{intra} and ΔG_{solv} are the intramolecular strain and desolvation penalty for each ligand upon binding. The conformational entropies (S_{conf}) in solution were computed from the probabilities (P_i) assuming a Boltzmann distribution (eq 2), where k_B is the Boltzmann constant. In the bound state, it was assumed that there was only one conformation accessible to each ligand; its conformational entropy is therefore zero. Thus, $-T\Delta S_{\text{conf}}$ is the ligand conformational entropy penalty, multiplied by the temperature to convert it into energy. E_{VDW} and E_{Elect} are the protein–ligand intermolecular van der Waals and electrostatic interaction energies, respectively. E_{PTN} , the protein

energy in the bound state, describes the protein deformation or strain imposed by each ligand. E_{GB} is the solvent shielding of protein–ligand electrostatic interactions estimated by the GB model. This term, generally not included in the scoring as it has no significant impact on the MM-GB/SA results,⁴ is necessary here as it depends on ϵ_{in} . Finally, $\Delta G_{\text{solv}}^{\text{ptn}}$ is the protein desolvation term calculated by the continuum model. The final ranking is obtained by calculating relative binding energies, $\Delta\Delta G_{\text{bind}}$, using the top-scoring inhibitor as a reference.

$$\Delta G_{\text{bind}} = \Delta E_{\text{intra}} + \Delta G_{\text{solv}} - T\Delta S_{\text{conf}} + E_{\text{VDW}} + E_{\text{Elect}} + E_{\text{GB}} + E_{\text{PTN}} + \Delta G_{\text{solv}}^{\text{ptn}} \quad (1)$$

$$S_{\text{conf}} = -k_B \sum_{i=1}^n P_i \ln P_i \quad (2)$$

The conformational search for the inhibitors in the unbound state and energy minimization for the complexes were performed with BOSS and MCPRO,²² respectively, instead of MacroModel.¹³ This was done with the purpose of facilitating the implementation of the variable dielectric approach. The Z matrices for the complexes obtained from the docking calculations, and ligands in the unbound state were prepared using the pepz program.²² The proteins were considered in their entirety. The protonation states of histidine side chains were assigned with the assistance of the software PROPKA 2.0.²³ Charge neutrality for the protein systems in MCPRO was imposed by assigning normal protonation states at physiological pH to basic and acidic residues near the active site and making the adjustments for neutrality to the most distant residues. The OPLS-AA force field was used for the protein.^{12a} The energetics for the ligands were represented with the OPLS/CM1A force field.²⁴ The CM1A atomic charges for the neutral ligands were scaled by 1.14.²⁵ The GB/SA solvation model implemented in BOSS and MCPRO was used.²⁶

The Variable Dielectric Formalism and Implementation. The focus here is to obtain a set of dielectric constants based on residue types that improves MM-GB/SA scoring. Protein intramolecular electrostatic interactions were not considered in the optimization process as the set of optimal dielectric constants for each residue–residue pair is not necessarily the best that can be obtained to improve the description of ligand–residue electrostatic interactions. In addition, the E_{PTN} values within a congeneric series generally fall in a very narrow range due to the constraints applied in the energy minimizations for the complexes.⁴ As this term behaves almost like a constant, it was excluded from the scoring equation used to develop the variable dielectric formalism. Although $\Delta G_{\text{solv}}^{\text{ptn}}$, like E_{PTN} , depends on ϵ_{in} , it was also excluded from the scoring equation, as it is usually very noisy and deteriorates the correlation with experimental data.^{4b,c} The energy minimizations for the complexes, however, are performed using GB/SA.

Equations 3 and 4 describe the protein–ligand electrostatic interactions and the solvent shielding estimated by the GB model, respectively, where q_i and q_j are charges on atom i belonging to the protein and j belonging to ligand, r_{ij} is their distance, ϵ_{in} is the internal dielectric constant, ϵ_{solv} is the dielectric constant in water, and α_{ij} is the geometric average of the Born radii α_i and α_j .

$$E_{\text{elect}} = \sum_{i,j} \frac{q_i q_j}{\epsilon_{\text{in}} r_{ij}} \quad (3)$$

Table 1. Optimal Set of Residue-Based Dielectric Constants ($\epsilon_{in(k)}$) for the Targets Individually and Best Overall

res type	CDK2		FactorXa		PDE10A		p38_u		optimal
Ser	20	1	1	1	20	1	20	20	1
Thr	20	2	20	2	1	2	1	1	4
Asn	1	1	20	1	20	8	1	4	2
Gln	1	1	1	20	4	2	1	1	1
His	20	1	20	20	20	1	1	20	4
Lys	2	2	20	1	20	20	1	1	2
Arg	1	1	1	1	2	1	20	20	20
Asp	1	8	20	20	2	20	20	20	2
Glu	20	20	1	1	1	1	20	20	8
Other	1	1	4	8	20	4	20	2	4
	$R^2 = 0.56$	PI = 0.76	$R^2 = 0.68$	PI = 0.90	$R^2 = 0.53$	PI = 0.71	$R^2 = 0.58$	PI = 0.75	

$$E_{GB} = - \left(\frac{1}{\epsilon_{in}} - \frac{1}{\epsilon_{sol}} \right) \sum_{i,j} \frac{q_i q_j}{\sqrt{r_{ij}^2 + \alpha_{ij}^2 \exp\left(\frac{r_{ij}^2}{4\alpha_{ij}^2}\right)}} \quad (4)$$

If ϵ_{in} adopts different values depending on the residue type k interacting with the ligand, eqs 3 and 4 have to be rewritten as shown in eqs 5 and 6. The dielectric constant $\epsilon_{in(k)}$ is assigned for the interactions between all atoms i belonging to the residue type k and the ligand atoms j . The side chains for all polar (Ser, Thr, Asn, Gln) and ionizable residues (His, Lys, Arg, Asp, Glu), which are expected to be more polarizing, were considered individually in the optimization. All remaining residue side chains and backbone atoms were bundled in a group called “other”. For simplicity, neutral and protonated states of the ionizable residues were not treated separately. The dielectric constant $\epsilon_{in(k)}$ for a given residue–ligand pair is the same whether the residue interacts with a neutral or charged ligand.

$$E_{elect} = \sum_k \frac{1}{\epsilon_{in(k)}} \sum_{i \in k, j} \frac{q_i q_j}{r_{ij}} \quad (5)$$

$$E_{GB} = - \sum_k \left(\frac{1}{\epsilon_{in(k)}} - \frac{1}{\epsilon_{sol}} \right) \sum_{i \in k, j} \frac{q_i q_j}{\sqrt{r_{ij}^2 + \alpha_{ij}^2 \exp\left(\frac{r_{ij}^2}{4\alpha_{ij}^2}\right)}} \quad (6)$$

Five different values (1, 2, 4, 8, and 20) were considered for every residue type (Ser, Thr, Asn, Gln, His, Lys, Arg, Asp, Glu, and other). Computational efficiency was achieved by obtaining the inner summation terms in eqs 5 and 6 for every residue k through one single point calculation per complex using the geometries obtained with $\epsilon_{in} = 1$. No cutoffs were applied to the protein–ligand electrostatic interactions. The inner summation terms were then combined with 5^{10} dielectric constant permutations outside of MCPRO to generate the MM-GB/SA scores corresponding to each set of E_{elect} and E_{GB} values. Their performance against the experimental data was judged on the basis of correlation coefficients (R^2) and predictive indices (PI).¹⁹ The latter is a measure of how accurate the predicted rank order is compared to the experiment, with -1 , 0 , and $+1$ meaning opposite, random, or perfect predictions, respectively. The set of dielectric constants that overall provided the best R^2 and PI values

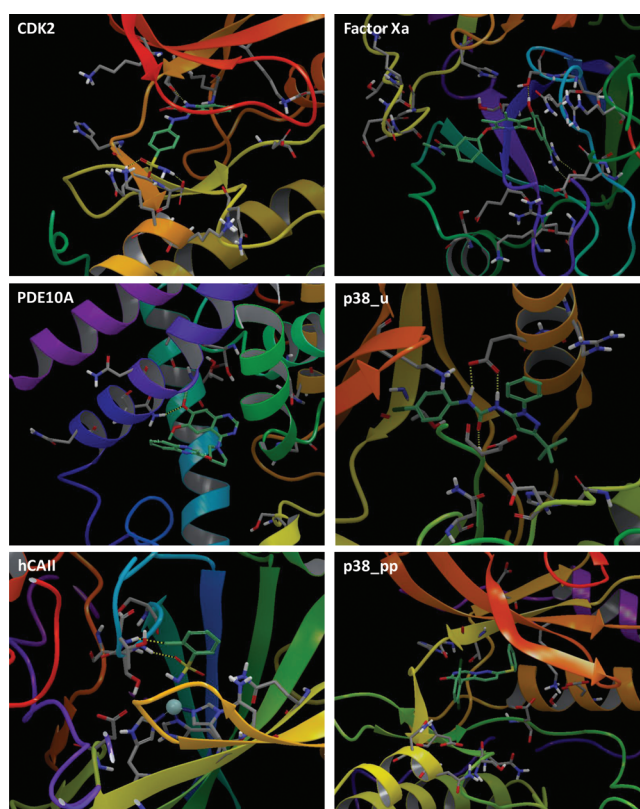


Figure 2. Ligands in their binding sites. Only the polar and ionizable residues within 6 Å from each ligand are shown for CDK2, Factor Xa, PDE10A, hCAII, and the p38 urea (p38_u) pyridazo-pyrimidinone (p38_pp) series.

for the training set targets and ligands was selected and then tested on two additional data sets.

RESULTS AND DISCUSSION

Residue-Based Dielectric Constant Optimization. In the optimization process, 5^{10} combinations of dielectric constants based on the residue type were generated for each system in the training set, CDK2, FactorXa, PEDE10A, and the p38 urea series (p38_u). The combinations that provided the MM-GB/SA scores with the best agreement with binding data based on R^2 and PI values for each target individually were set aside.

Table 2. MM-GB/SA Results Using Standard and Variable Dielectric Protein–Ligand Electrostatics

target	N^a	out ^b	R^2 STD ^c	R^2 NO ^d	R^2 VAR ^e	PI STD	PI NO	PI VAR	DR ^f STD	DR NO	DR VAR
CDK2	36	7	0.48	0.41	0.52	0.69	0.66	0.73	17.6	13.7	13.1
FactorXa	22	3	0.59	0.09	0.66	0.84	0.37	0.87	14.8	6.3	9.3
PDE10A	28	0	0.47	0.52	0.51	0.65	0.69	0.67	8.3	6.7	7.1
p38_u	36	2	0.33	0.56	0.55	0.46	0.67	0.71	20.9	21.0	21.9
hCAII	17	2	0.32	0.39	0.58	0.63	0.74	0.75	11.9	10.9	7.8
p38_pp	17	3	0.41	0.49	0.43	0.68	0.70	0.70	10.5	9.4	10.5

^aNumber of total compounds in each series. ^bNumber of outliers removed. ^cStandard dielectric protein–ligand electrostatics (STD). ^dNo protein–ligand electrostatics. E_{elect} and E_{GB} are removed from the scoring equation (NO). ^eVariable dielectric protein–ligand electrostatics (VAR). ^fScore dynamic range (DR) in kcal/mol.

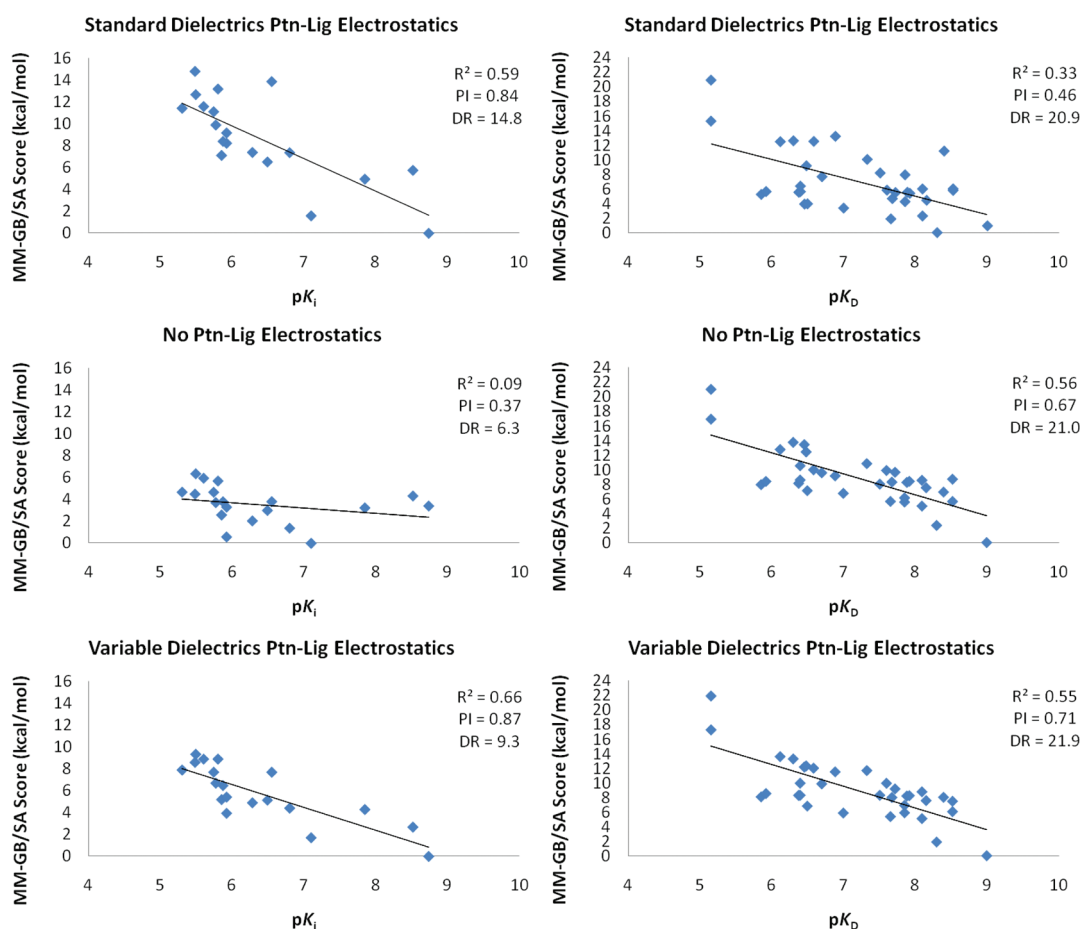


Figure 3. MM-GB/SA scoring versus experimental results using standard and variable dielectric protein–ligand electrostatics, and with the E_{elect} and E_{GB} terms removed for two systems, Factor Xa (left) and p38_u (right).

Table 1 illustrates the combinations for each target that maximized R^2 and PI separately. Although the combinations listed in Table 1 are the very best for each system according to each metric, there were few other solutions that provided just slightly worse results. The set of optimal dielectric constants for all targets simultaneously was derived by verifying the combination whose R^2 and PI values deviated the least from each individual's best (Table 1).

Figure 2 shows the polar and ionizable residues within 6 Å from the ligands depicted in Figure 1 for the systems that belong not only to the training set but also to the test set, hCAII and the p38 pyridazo-pyrimidinone (p38_pp) series. As electrostatic

interactions are very long-range, it is obvious that polar and ionizable residues beyond 6 Å are also relevant; their interactions with the ligands are included in E_{elect} and E_{GB} . Although the optimal set of dielectric constants is a function of the specific residue–ligand interactions in the training set and inaccuracies in the charges of the force field of choice, it is tempting to physically interpret the $\epsilon_{\text{in}(k)}$ values obtained. Here, Arg and Glu have the largest $\epsilon_{\text{in}(k)}$ values, 20 and 8, respectively, which seems reasonable since those residues, very flexible and polarizing, can be more easily shielded. The optimal value for Asp, with a shorter side chain and reduced flexibility when compared to Glu, is 2.

Interestingly, a much smaller dielectric constant was obtained for Lys in contrast to the one for Arg. In a study of residue density in proteins,²⁷ it became evident that in spite of evolutionary relatedness, Arg is more buried and more frequently involved in salt bridges, hydrogen bonds, and cationic–aromatic contacts. It is then plausible that a larger $\epsilon_{\text{in}(k)}$ value than that for Lys emerges for Arg in the optimization process. In the case of the polar and the ionizable His residues, found to be in its neutral state in the majority of the cases here, the $\epsilon_{\text{in}(k)}$ values are fairly small, ranging from 1 to 4. For three of the five residues, the dielectric constants seem to be correlated with the magnitude of the side chain dipole moments, with the $\epsilon_{\text{in}(k)}$ value for His > Asn > Ser. As for Gln and Thr, the dielectric constants obtained seem to be counterintuitive when compared to their closest analogs, Asn and Ser, respectively. The optimal $\epsilon_{\text{in}(k)}$ value for the more flexible side chain of the pair is actually smaller; this might be a function of the specific residue distributions for the targets in the training set. Finally, the $\epsilon_{\text{in}(k)}$ value of 4 for the set that contains all remaining side chains might be somewhat high because it also contains all backbone atoms.

Standard versus Variable Dielectric MM-GB/SA Scoring. It should be noted that the variable dielectric approach was unable to recover some outliers in the training set in the preliminary rounds of dielectric constant optimization. Those compounds had to be ultimately excluded before the optimization round that led to the $\epsilon_{\text{in}(k)}$ values in Table 1. The number of outliers removed, in both the training and test sets, is listed in Table 2. Except for CDK2, the outliers represent between 0% and 15% of the compounds in each series and are highlighted in the tables provided in the Supporting Information. The balance between the number of remaining binding data points in the training set (110) and the number of adjustable $\epsilon_{\text{in}(k)}$ values (10), a 11:1 ratio, seems appropriate and suggests that the set of dielectric constants obtained might be applied more widely. Table 2 illustrates the R^2 and PI values obtained using the standard and variable dielectric MM-GB/SA scoring procedures, the latter using the optimal $\epsilon_{\text{in}(k)}$ values in Table 1. The results where E_{elect} and E_{GB} are removed from the scoring equation are also included for comparison. They provide a baseline for the cases where the standard dielectric MM-GB/SA performs poorly. In those cases, the variable dielectric approach should at least outperform the scoring equation with E_{elect} and E_{GB} removed.

It is disturbing that the scoring equation with no protein–ligand electrostatics performs better than the standard MM-GB/SA in four out of the six systems (Table 2). Factor Xa, where E_{elect} and E_{GB} are critical, and CDK2, where they marginally improve the results, are the only exceptions. In Factor Xa, the most potent compounds of the series establish a key hydrogen bond with the catalytic triad Ser residue (Figure 2), which cannot be described without the electrostatic terms. Although it is difficult to pinpoint whether the deterioration of the correlation with experimental data when E_{elect} and E_{GB} are introduced has its origin in the force field charges, the lack of explicit polarization, or the lack of dynamical screening due to the use of a single structure for the complex, it is clear that the variable dielectric approach improves the description of protein–ligand electrostatics. It performs better than standard dielectrics in all cases, including the two systems in the test set, hCAII and p38_pp. Figure 3 illustrates the results for Factor Xa and p38_u using the different MM-GB/SA scoring approaches. Plots for all systems can be found in the Supporting Information.

It is also reassuring that the performance of the variable dielectric approach is at least equivalent or superior to the scoring equation with E_{elect} and E_{GB} removed. A scoring method that describes protein–ligand electrostatic interactions properly as well as the fine balance with the desolvation penalty process is highly desirable since the introduction of polarity, which drives the compound to a better property space,²⁸ often kills or attenuates binding affinity. One should note that the variable dielectric approach will not necessarily provide significant improvements over the standard electrostatic treatment for all cases. This is illustrated in Table 2 for a couple of systems, PDE10A and p38_pp. In those instances, the residues around the ligand are mostly nonpolar, and the electrostatic interactions between them are not appreciably large (Figure 2). Specifically, the combined E_{elect} and E_{GB} terms for all ligands in the congeneric series obtained with $\epsilon_{\text{in}} = 1$ range from -1.9 to -0.2 kcal/mol and -3.6 to 0.0 kcal/mol for PDE10A and p38_pp, respectively. In other words, protein–ligand electrostatic interactions play a minor role in the rank-ordering for the PDE10A and p38_pp, and the attenuation or exclusion of E_{elect} and E_{GB} has no impact on the results. This contrasts with p38_u, for example, with a much wider range (-11.8 to $+1.9$ kcal/mol) for the combined E_{elect} and E_{GB} terms. The protein conformation (DFG out) for p38, the chemical series, and binding mode are very different, with the urea group establishing hydrogen bonds with a Glu residue and a backbone NH (Figure 2). In this case, the variable dielectric approach greatly helps since E_{elect} and E_{GB} are relevant. Finally, Table 2 indicates that the variable dielectric approach reduces the score dynamic range (DR) in four of the six systems, although no attempt has been made to improve DR in the optimization process. This supports the view that the exaggerated protein–ligand electrostatic interactions due to the lack of shielding effects in the standard MM-GB/SA model are indeed a key factor in its wide scoring spread.

■ ASSOCIATED CONTENT

S Supporting Information. Tables with all derivatives and binding data and plots of the multiple MM-GB/SA approaches against the experiment for all systems are available. This material is available free of charge via the Internet at <http://pubs.acs.org>.

■ AUTHOR INFORMATION

Corresponding Author

*Phone: (860) 686-2915. E-mail: cristiano.guimaraes@pfizer.com.

■ ACKNOWLEDGMENT

The authors thank the Computational Sciences Center of Emphasis at Pfizer and the National Institutes of Health (GM32136 and AI44616) for support.

■ REFERENCES

- (1) Barril, X.; Gelpí, J. L.; López, J. M.; Orozco, M.; Luque, F. J. How accurate can molecular dynamics/linear response and Poisson–Boltzmann/solvent accessible surface calculations be for predicting relative binding affinities? Acetylcholinesterase huprine inhibitors as a test case. *Theor. Chem. Acc.* **2001**, *106*, 2–9.
- (2) Kuhn, B.; Kollman, P. A. Binding of a diverse set of ligands to avidin and streptavidin: An accurate quantitative prediction of their

relative affinities by a combination of molecular mechanics and continuum solvent models. *J. Med. Chem.* **2000**, *43*, 3786–3791.

(3) (a) Hou, T.; Wang, J.; Li, Y.; Wang, W. Assessing the performance of the MM/PBSA and MM/GBSA methods. 1. The accuracy of binding free energy calculations based on molecular dynamics simulations. *J. Chem. Inf. Model.* **2011**, *51*, 69–82. (b) Haider, M. K.; Bertrand, H.-O.; Hubbard, R. E. Predicting fragment binding poses using a combined MCSS MM-GBSA approach. *J. Chem. Inf. Model.* **2011**, *51*, 1092–1105. (c) Lee, M. R.; Sun, Y. Improving docking accuracy through molecular mechanics generalized Born optimization and scoring. *J. Chem. Theory Comput.* **2007**, *3*, 1106–1119. (d) Foloppe, N.; Hubbard, R. Towards predictive ligand design with free-energy based computational methods?. *Curr. Med. Chem.* **2006**, *13*, 3583–3608. (e) Huang, N.; Kalyanaraman, C.; Irwin, J. J.; Jacobson, M. P. Physics-based scoring of protein-ligand complexes: Enrichment of known inhibitors in large-scale virtual screening. *J. Chem. Inf. Model.* **2006**, *46*, 243–253. (f) Huang, N.; Kalyanaraman, C.; Bernacki, K.; Jacobson, M. P. Molecular mechanics methods for predicting protein–ligand binding. *Phys. Chem. Chem. Phys.* **2006**, *8*, 5166–5177. (g) Lyne, P. D.; Lamb, M. L.; Saeh, J. C. Accurate prediction of the relative potencies of members of a series of kinase inhibitors using molecular docking and MM-GBSA scoring. *J. Med. Chem.* **2006**, *49*, 4805–4808.

(4) (a) Guimarães, C. R. W. A direct comparison of the MM-GB/SA scoring procedure and free-energy perturbation calculations using Carbonic Anhydrase as a test case: Strengths and pitfalls of each approach. *J. Chem. Theory Comput.* **2011**, *7*, 2296–2306. (b) Guimarães, C. R. W.; Mathiowetz, A. M. Addressing limitations with the MM-GB/SA scoring procedure using the WaterMap method and free-energy perturbation calculations. *J. Chem. Inf. Model.* **2010**, *50*, 547–559. (c) Guimarães, C. R. W.; Cardozo, M. MM-GB/SA rescoring of docking poses in structure-based lead optimization. *J. Chem. Inf. Model.* **2008**, *48*, 958–970.

(5) (a) Guimarães, C. R. W.; Boger, D. L.; Jorgensen, W. L. Elucidation of fatty acid amide hydrolase inhibition by potent α -ketoheterocycle derivatives from Monte Carlo simulations. *J. Am. Chem. Soc.* **2005**, *127*, 17377–17384. (b) Simonson, T.; Georgios, A.; Karplus, M. Free energy simulations come of age: Protein-ligand recognition. *Acc. Chem. Res.* **2002**, *35*, 430–437. (c) Jorgensen, W. L. Free energy changes in solution. In *Encyclopedia of Computational Chemistry*; Schleyer, P. v. R., Ed.; Wiley: New York, 1998; Vol. 2, pp 1061–1070. (d) Kollman, P. A. Free energy calculations: Applications to chemical and biochemical phenomena. *Chem. Rev.* **1993**, *93*, 2395–2417. (e) Jorgensen, W. L. Free energy calculations: A breakthrough for modeling organic chemistry in solution. *Acc. Chem. Res.* **1989**, *22*, 184–189.

(6) 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.

(7) Abel, R.; Salam, N. K.; Shelley, J.; Farid, R.; Friesner, R. A.; Sherman, W. Contribution of explicit solvent effects to the binding affinity of small-molecule inhibitors in blood coagulation factor serine proteases. *ChemMedChem* **2011**, *6*, 1049–1066.

(8) Schutz, C. N.; Warshel, A. What are the dielectric “constants” of proteins and how to validate electrostatic models?. *Proteins* **2001**, *44*, 400–417.

(9) Sham, Y. Y.; Muegge, I.; Warshel, A. The effect of protein relaxation on charge–charge interactions and dielectric constants of proteins. *Biophys. J.* **1998**, *74*, 1744–1753.

(10) Muegge, I.; Schweins, T.; Langen, R.; Warshel, A. Electrostatic control of gtp and gdp binding in the oncoprotein p21 ras. *Structure* **1996**, *4*, 475–489.

(11) Zhu, K.; Shirts, M. R.; Friesner, R. A. Improved methods for side chain and loop predictions via the protein local optimization program: Variable dielectric model for implicitly improving the treatment of polarization effects. *J. Chem. Theory Comput.* **2007**, *3*, 2108–2119.

(12) (a) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. Development and testing of OPLS all-atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc.* **1996**, *118*, 11225–11236. (b) Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. J. Evaluation and reparametrization of the OPLS-AA force

field for proteins via comparison with accurate quantum chemical calculations on peptides. *J. Phys. Chem. B* **2001**, *105*, 6474–6487.

(13) *MacroModel*, version 9.0; Schrödinger, LLC: New York, 2005.

(14) (a) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. Glide: A new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.* **2004**, *47*, 1739–1749. (b) Friesner, R. A.; Murphy, R. B.; Repasky, M. P.; Frye, L. L.; Greenwood, J. R.; Halgren, T. A.; Sanschagrin, P. C.; Mainz, D. T. Extra precision Glide: docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes. *J. Med. Chem.* **2006**, *49*, 6177–6196.

(15) Bramson, H. N.; Corona, J.; Davis, S. T.; Dickerson, S. H.; Edelstein, M.; Frye, S. V.; Gampe, R. T.; Harris, J. P. A.; Hassell, A.; Holmes, W. D.; Hunter, R. N.; Lackey, K. E.; Lovejoy, B.; Luzzio, M. J.; Montana, V.; Rocque, W. J.; Rusnak, D.; Shewchuk, L.; Veal, J. M.; Walker, D. H.; Kuyper, L. F. Oxindole-based inhibitors of Cyclin-Dependent Kinase 2 (CDK2): Design, synthesis, enzymatic activities, and X-ray crystallographic analysis. *J. Med. Chem.* **2001**, *44*, 4339–4358.

(16) (a) Phillips, G.; Davey, D. D.; Eagen, K. A.; Koovakkat, S. K.; Liang, A.; Ng, H. P.; Pinkerton, M.; Trinh, L.; Whitlow, M.; Beatty, A. M.; Morrissey, M. M. Design, synthesis, and activity of 2,6-diphenoxypyridine-derived Factor Xa inhibitors. *J. Med. Chem.* **1999**, *42*, 1749–1756. (b) Phillips, G.; Guilford, W. J.; Buckman, B. O.; Davey, D. D.; Eagen, K. A.; Koovakkat, S.; Liang, A.; McCarrick, M.; Mohan, R.; Ng, H. P.; Pinkerton, M.; Subramanyam, B.; Ho, E.; Trinh, L.; Whitlow, M.; Wu, S.; Xu, W.; Morrissey, M. M. Design, synthesis, and activity of a novel series of Factor Xa inhibitors: Optimization of arylamidine groups. *J. Med. Chem.* **2002**, *45*, 2484–2493.

(17) (a) Verhoest, P. R.; Chapin, D. S.; Corman, M.; Fonseca, K.; Harms, J. F.; Hou, X.; Marr, E. S.; Menniti, F. S.; Nelson, F.; O'Connor, R.; Pandit, J.; Proulx-LaFrance, C.; Schmidt, A. W.; Schmidt, C. J.; Suiciak, J. A.; Liras, S. Discovery of a novel class of Phosphodiesterase 10A inhibitors and identification of clinical candidate 2-[4-(1-Methyl-4-pyridin-4-yl-1H-pyrazol-3-yl)-phenoxy-methyl]-quinoline (PF-2545920) for the treatment of schizophrenia. *J. Med. Chem.* **2009**, *52*, 5188–5196. (b) Chappie, T. A.; Humphrey, J. M.; Allen, M. P.; Estep, K. G.; Fox, C. B.; Lebel, L. A.; Liras, S.; Marr, E. S.; Menniti, F. S.; Pandit, J.; Schmidt, C. J.; Tu, M.; Williams, R. D.; Yang, F. Y. Discovery of a series of 6,7-dimethoxy-4-pyrrolidylquinazoline PDE10A inhibitors. *J. Med. Chem.* **2007**, *50*, 182–185.

(18) Regan, J.; Breitfelder, S.; Cirillo, P.; Gilmore, T.; Graham, A. G.; Hickey, E.; Klaus, B.; Madwed, J.; Moriak, M.; Moss, N.; Pargellis, C.; Pav, S.; Proto, A.; Swinamer, A.; Tong, L.; Torcellini, C. Pyrazole urea-based inhibitors of p38 MAP kinase: From lead compound to clinical candidate. *J. Med. Chem.* **2002**, *45*, 2994–3008.

(19) Pearlman, D. A.; Charifson, P. S. Are free energy calculations useful in practice? A comparison with rapid scoring functions for the p38 MAP kinase protein system. *J. Med. Chem.* **2001**, *44*, 3417–3423.

(20) Scott, A. D.; Phillips, C.; Alex, A.; Flocco, M.; Bent, A.; Randall, A.; O'Brien, R.; Damian, L.; Jones, L. H. Thermodynamic optimization in drug discovery: A case study using Carbonic Anhydrase inhibitors. *Chem. Med. Chem.* **2009**, *4*, 1985–1989.

(21) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. Semianalytical treatment of solvation for molecular mechanics and dynamics. *J. Am. Chem. Soc.* **1990**, *112*, 6127–6129.

(22) Jorgensen, W. L.; Tirado-Rives, J. Molecular modeling of organic and biomolecular systems using BOSS and MCPRO. *J. Comput. Chem.* **2005**, *26*, 1689–1700.

(23) Bas, D. C.; Rogers, D. M.; Jensen, J. H. Very fast prediction and rationalization of pK_a values for protein–ligand complexes. *Proteins* **2008**, *73*, 765–783.

(24) Jorgensen, W. L.; Tirado-Rives, J. Potential energy functions for atomic-level simulations of water and organic and biomolecular systems. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6665–6670.

(25) Udier-Blagovic, M.; Morales De Tirado, P.; Pearlman, S. A.; Jorgensen, W. L. Accuracy of free energies of hydration from CM1 and CM3 atomic charges. *J. Comput. Chem.* **2004**, *25*, 1322–1332.

(26) 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.

(27) Baud, F.; Karlin, S. Measures of residue density in protein structures. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 12494–12499.

(28) (a) Hughes, J. D.; Blagg, J.; Price, D. A.; Bailey, S.; DeCrescenzo, G. A.; Devraj, R.; Ellsworth, E.; Fobian, Y. M.; Gibbs, M. E.; Gilles, R. W.; Greene, N.; Huang, E.; Krieger-Burke, T.; Loesel, J.; Wager, T.; Whiteley, L.; Zhang, Y. Physiochemical drug properties associated with in vivo toxicological outcomes. *Bioorg. Med. Chem. Lett.* **2008**, *18*, 4872–4875.
(b) Wager, T. T.; Hou, X.; Verhoest, P. R.; Villalobos, A. Moving beyond rules: The development of a central nervous system multiparameter optimization (CNS MPO) approach to enable alignment of druglike properties. *ACS Chem. Neurosci.* **2010**, *1*, 420–434.

Performance of the van der Waals Density Functional VV10 and (hybrid)GGA Variants for Thermochemistry and Noncovalent Interactions

Waldemar Hujo and Stefan Grimme*

Mulliken Center for Theoretical Chemistry, Institut für Physikalische und Theoretische Chemie, Universität Bonn, Beringstr. 4, D-53115 Bonn, Germany

S Supporting Information

ABSTRACT: The nonlocal van der Waals density functional VV10 (Vydrov, O. A.; Van Voorhis, T. J. *Chem. Phys.* **2010**, *133*, 244103) is tested for the thermochemical properties of 1200+ atoms and molecules in the GMTKN30 database in order to assess its global accuracy. Five GGA and hybrid functionals in unmodified form are augmented by the nonlocal (NL) part of the VV10 functional (one parameter adjusted). The addition of the NL dispersion energy definitely improves the results of all tested functionals. On the basis of little empiricism and basic physical insight, DFT-NL can be recommended as a fully electronic, robust electronic structure method.

INTRODUCTION

Accounting for the missing van der Waals (vdW, dispersion) forces^{1,2} in standard Kohn–Sham Density Functional Theory (DFT) has become essential in many studies of chemical and physical electronic structure problems. Over the past decades, significant development (for an overview see ref 3) has been followed with great success yielding methods which are not only efficient but also accurate in describing a broad range of chemical and biological systems.^{4–6} Common approaches use atom pairwise additive schemes,⁷ extensive density functional parametrization,⁸ or effective one-electron potentials⁹ (for mixed density-based/atom pairwise schemes, see refs 10–13).

A different route is followed by the van der Waals Density Functionals (vdW-DF).^{14–16} These methods use as input to compute the Non-Local (NL) dispersion contribution only the electron density. The recently developed vdW functional VV10 of Vydrov and Van Voorhis¹⁷ currently seems to be the most promising candidate for a general and accurate method. For molecules involving weak hydrogen bonds, which are significantly influenced by dispersion interactions, the VV10 functional (and to a lesser degree also the vdW-DF2¹⁶) showed remarkably good performance in a recent study conducted in our group.¹⁸ One achievement of VV10, besides its simplicity, is a consistent description of interaction energies and equilibrium distances. A unique feature of the vdW-DFs is that the change of the electron density due to dispersion effects is accounted for, if implemented self-consistently. Furthermore, these methods—based on first principles—use only one (vdW-DF2) or two (VV10) global empirical parameters, which are necessary for a seamless and consistent connection with the necessary semilocal DF parts.

So far, VV10 has mainly been tested for noncovalent interactions (NCI), and it is not really known how it performs for more complicated energetic properties (only six atomization energies have been investigated in ref 17). Because the long-range dispersion energy can be considered as becoming part of the “normal” electron correlation,^{19,20} dispersion corrections also

influence the description of thermochemistry where typically strong bonds are formed. A very convenient and unbiased way to assess the “global” accuracy of VV10 is using the GMTKN30⁶ database that was recently developed in our group. This benchmark covers 30 subsets related to general main group thermochemistry, kinetics, and NCI. In total, it encompasses 1218 single-point calculations and 841 data points (relative energies); therefore, it turned out to be ideal for the evaluation and development of DFT methods. The subsets of the GMTKN30 database are divided into three major sections, which are discussed below for analysis purposes.

We utilize the GMTKN30 database to evaluate the original VV10 functional and its nonlocal part with other GGA and hybrid density functionals. To put this into broader perspective, we compare its performance with plain DFT and the nondensity-based DFT-D3^{7,20} method, which is already known to improve the accuracy of standard DFT significantly.⁵ Besides the fundamental question of practical applicability of VV10-type functionals, one important aspect of the present work is whether the NL part can be added to standard functionals, and whether double-counting effects of correlation between the semilocal and NL components play a role.

METHODS AND COMPUTATIONAL DETAILS

The GMTKN30 benchmark database used in our study employs a weighted total mean absolute deviation of relative energies (WTMAD, for details, see ref 6) as a measure of accuracy, which is also used here. As usual, fixed molecular structures are used. The original vdW functional VV10 as proposed by Vydrov and Van Voorhis¹⁷ and in addition five new variants of it are evaluated in detail. In its original formulation, VV10 uses the underlying GGA functional rPW86PBE.^{21,22} We combine the GGA density functionals BLYP^{23,24} and revPBE^{22,25}

Received: September 15, 2011

Published: October 25, 2011

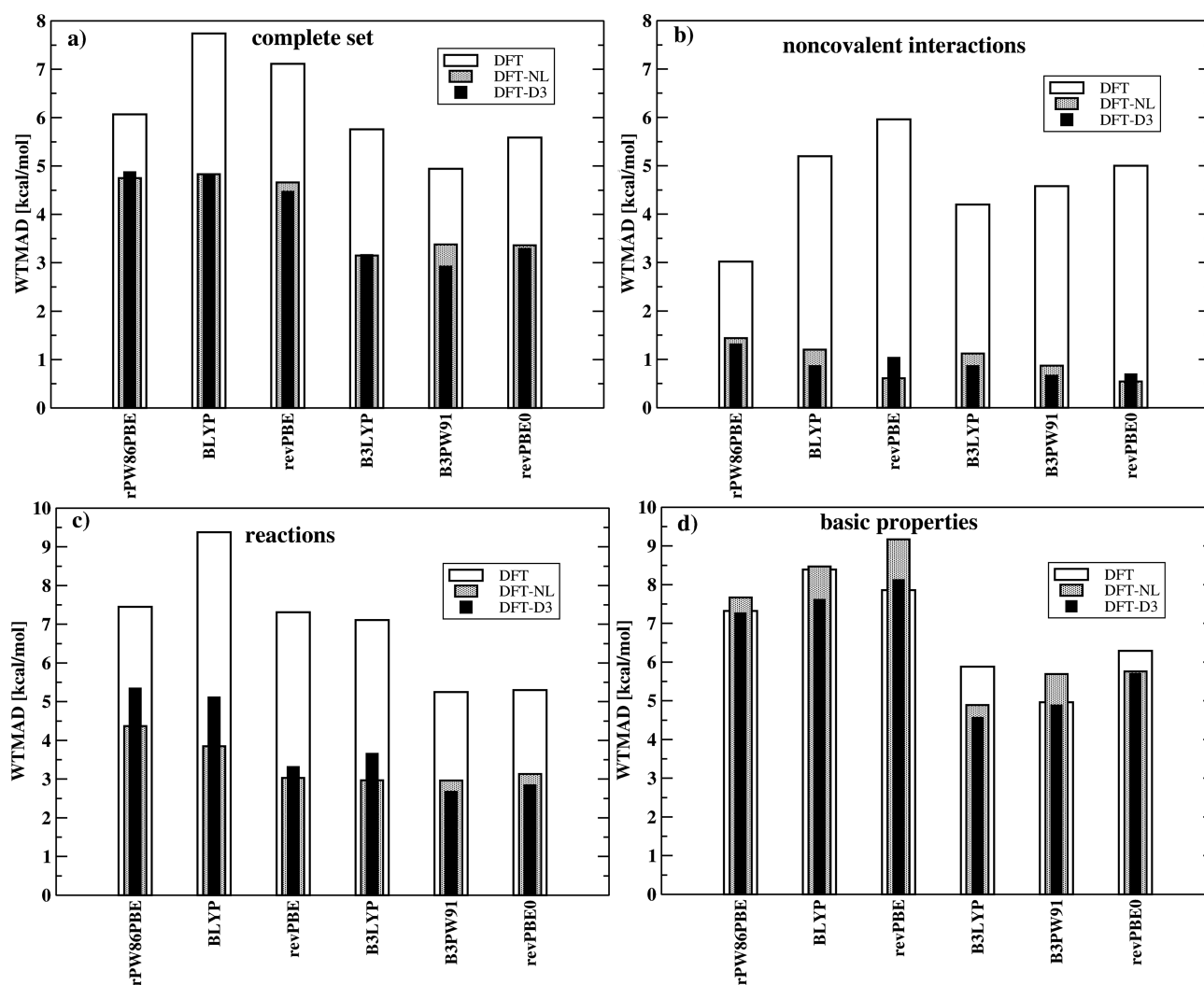


Figure 1. The weighted total mean absolute deviation (WTMAD) for the entire GMTKN30 database (a) for the three GGA functionals rPW86PBE, BLYP, and revPBE as well as the three global hybrids B3LYP, B3PW91, and revPBE0 are compared with the values for the corresponding DFT-D3 and DFT-NL augmented versions. Plots b–d show the corresponding results for its subsections.

and the global hybrids B3LYP,^{26,27} B3PW91,²⁶ and revPBE0^{25,28,29} with the NL part of VV10 and fit the latter to the S22³⁰ set of NCI energies by adjusting only one parameter. The GGA parts are not modified as in DFT-D3, which facilitates comparison and allows a convenient implementation in existing program packages. The choice of (hybrid)GGAs has been made without any special intention (except for B3LYP, which is the most widely used functional), and we do not expect this to have any effect on our basic conclusions.

The total exchange–correlation (XC) energy of the original VV10 is defined in eq 1. It is composed of the refitted Perdew–Wang^{21,31} exchange (rPW86) with PBE²² correlation and an NL term, which covers the (mainly) long-range dispersive energy:

$$E_{xc}^{VV10} = E_{xc}^{rPW86PBE} - NL \\ = E_x^{rPW86} + E_c^{PBE} + E_{c-NL}^{VV10} \quad (1)$$

The NL term is given by the following double-integral:

$$E_{c-NL}^{VV10} = \int dr \rho(r) \left[\beta + \frac{1}{2} \int dr' \rho(r') \phi(r, r') \right] \quad (2)$$

where ρ is the total electron density, and the definition of the kernel $\phi(r, r')$ and β is as follows (in atomic units):

$$\phi(r, r') = -\frac{3}{2gg'(g+g')}g = \omega_0(r)R^2 + \kappa(r), \\ g' = \omega_0(r')R^2 + \kappa(r'), R = |r - r'|, \\ \omega_0(r) = \sqrt{C \left| \frac{\nabla \rho(r)}{\rho(r)} \right|^4 + \frac{4\pi}{3}\rho(r)}, \\ \kappa(r) = b \frac{3\pi}{2} \left[\frac{\rho(r)}{9\pi} \right]^{1/6}, \beta = \frac{1}{32} \left[\frac{3}{b^2} \right]^{3/4} \quad (3)$$

In the original definition, the short-range attenuation parameter b appearing in κ and β was fitted to the S22 set³⁰ of NCIs ($b = 5.9$ for rPW86PBE). We also used the S22 set to obtain optimum values of b for the other five functionals. The other parameter $C = 0.0093$ in E_{c-NL}^{VV10} , which determines the long-range behavior, was always set to its original value (as recommended for GGAs¹⁷) because initial optimization tests lead to only minor improvements. Note that the estimated accuracy for asymptotic

molecular dispersion coefficients is slightly better for D3 than for VV10 (mean relative errors of 5% and 9%, respectively^{7,17}). We obtain optimum values for b (in atomic units: 4.0 (BLYP), 3.7 (revPBE), 4.8 (B3LYP), 4.5 (B3PW91), and 4.3 (revPBE0), respectively) which are all slightly smaller than that for rPW86PBE. For convenience, we will abbreviate the combination of the VV10 nonlocal correlation part (with the above optimized b values) with a particular density functional DF as “DF-NL” (e.g., B3LYP-NL), which resembles the nomenclature in DFT-D3 where the corresponding method is named “B3LYP-D3”.

The sensitivity of the results on the parameter b is not very pronounced for changes of about $\pm 10\%$, as shown in the Supporting Information. We also tried to obtain a general formula for the parameter b for hybrid density functionals depending on the amount of Fock-exchange admixture. For this purpose, we also added the E_{c-NL}^{VV10} term to Hartree–Fock (HF) and revPBE38 (i.e., revPBE with a larger Fock-exchange admixture of $3/8 = 37.5\%$ instead of 0% in revPBE and 25% in revPBE0). For the two methods HF-NL ($b = 3.9$) and revPBE38-NL ($b = 4.7$), we computed S22 data and obtained very reasonable MAD values of 0.47 and 0.26 kcal/mol, respectively. However, we found no simple relation between the optimum value of b and the amount of Fock exchange included and hence suggest optimization of b for every new density functional individually. The HF-NL and revPBE38-NL methods are not discussed further.

In our current implementation,¹⁸ a standard SCF run with the semilocal (hybrid) potential is performed first, and the E_{c-NL}^{VV10} term based on the converged density is added. This non-self-consistent procedure is quite accurate^{14–18,32–35} and also consistent with the DFT-D3 treatment.

The simplest way of providing an asymptotically correct $-R^{-6}$ dependence of the dispersion energy on the intermolecular distance R is the DFT-D approach (for a recent review see, e.g., ref 3). Opposed to vdW-DFs, the method provides a nonelectronic dispersion energy without any significant additional computational cost. Recently, we tested the new DFT-D3⁷ with Becke’s and Johnson’s (BJ) rational damping function.^{10,36,37} This so-called BJ-damping²⁰ leads to a constant contribution of dispersion to the total correlation energy from each spatially close pair of atoms, which resembles the seamless description in VV10, which also includes a finite $R \rightarrow 0$ component.¹⁷ In the following, we use the term “DFT-D” as a synonym for “DFT-D3(BJ)” (and “-D3” for “-D3(BJ)”).

All DFT calculations were carried out with a locally modified version of the ORCA³⁸ program. The computations employ the resolution of the identity method (RI)³⁹ and the huge def2-QZVP AO basis set, which provides results quite close to the basis set limit (for details, see refs 5 and 6). For the numerical quadrature in the SCF calculations, we chose *Grid6*, and for the NL calculations, we used *Grid4* in the ORCA program. Basis set superposition errors are negligible for (hybrid)GGA computations with such a large basis set,⁴⁰ and consequently, counterpoise corrections are not applied.

RESULTS AND DISCUSSION

We first discuss results of the three methods DFT, DFT-D3, and DFT-NL for the entire GMTKN30 database and its three subsections (for a detailed description, see ref 5). The deviations for individual sets are shown then for the example of B3LYP only. We end our discussion with the presentation of results for NCI in S22 and S66⁴¹ sets. More detailed statistical information can be found in the Supporting Information.

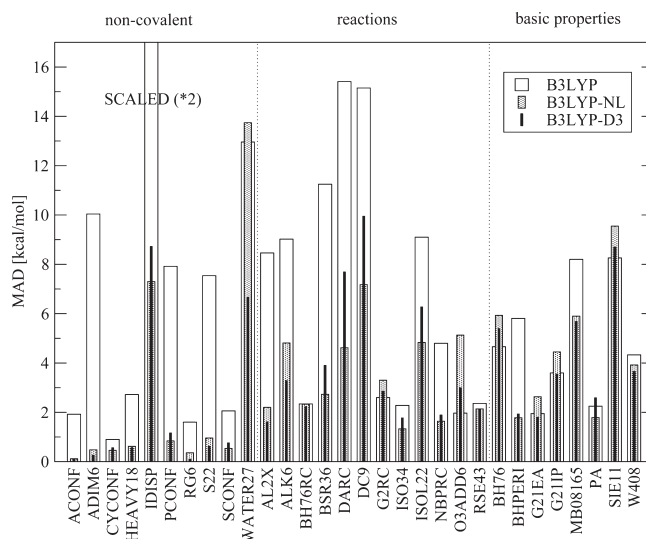


Figure 2. MAD values for all 30 sets of the GMTKN30 database are shown for B3LYP, B3LYP-NL, and B3LYP-D3. For better comparison, the NCI sets (left part) were scaled by a factor of 2.

It is clear from our recent studies^{5–7,20} that dispersion corrections to standard DFs significantly improve the accuracy for almost all chemically relevant properties (often by several kilocalories per mole). As can be seen in Figure 1, this holds true for DFT-D3 but, importantly, also for DFT-NL. This is a very fundamental result, as it shows that a slightly adjusted NL term together with a standard density functional treatment can be used not only for NCI (for which it was originally developed) but also for general thermochemistry. We are very satisfied to note that two fairly different dispersion corrections (D3 and NL) that, however, try to describe the same physical effects both perform very well for six different density functionals.

For the complete set, the WTMADs for uncorrected GGAs and hybrids are 6–8 kcal/mol and 5–6 kcal/mol, respectively. This is reduced to about 4.8 kcal/mol with D3 and NL corrections for all three GGAs. The hybrids benefit similarly and the WTMADs are reduced to around 3 kcal/mol. Except for B3PW91, for which the D3 correction seems to be superior, the performance of D3 and NL is very similar.

Dispersion corrections for DFT were originally designed with a focus on weakly (noncovalently) bound complexes, and for completeness we discuss the results for this NCI subset of GMTKN30 briefly (see Figure 1b, results for the S22³⁰ and S66⁴¹ sets are discussed separately below). As expected, the improvements by D3 and NL corrections are largest here. DFT-D3 performs slightly better for four out of the six functionals, but again, the accuracy of both approaches is rather similar. The resulting average errors of about 1 kcal/mol or better are highly satisfactory, keeping in mind that this part contains difficult conformational problems as well as large systems (e.g., $(H_2O)_{20}$ clusters with dissociation energies of about 200 kcal/mol).

Although not especially designed for thermochemical applications, D3/NL corrected functionals provide very encouraging results for important chemical reaction energies and improve plain DFT significantly (see Figure 1c). This is a main result of the present work and attributed to a significant contribution of relatively short-ranged intramolecular dispersion energies to the considered chemical reactions. Note that such thermochemical

data have not been used in the fitting of the empirical parameters contained in D3/NL. The chemical reactions in this section of GMTKN30 do not contain systems obviously dominated by “classical” long-range dispersion effects. Although the largest corrections are obtained for large (dense) molecule reactions (e.g., ISOL, DARC or BSR36 sets) where important dispersion effects are understandable, we also note improvements for sets containing relatively small molecules (e.g., ISO34). For a further discussion, see below and Figure 2. The WTMADs for reactions are almost halved in all cases. DFT-NL seems to outperform DFT-D3 here, as it provides better results for four functionals. This seems to be clear as the change of electronic structure during a chemical reaction is captured better by a density-based method than by an atom-pairwise approach. Compared to other more modern DFs (like M06-2X, PW6B95, or double-hybrids, see ref 5), the B3LYP-NL method provides a slightly worse but competitive accuracy. Note also the very good results for the “inexpensive” (GGA based) revPBE-NL functional. The better performance of the four (out of six) DFT-NL functionals compared to DFT-D3 for reaction energies results basically from a better description of bond separation reactions of saturated hydrocarbons (BSR36), Diels–Alder reactions (DARC), and to some extent also dimerization energies of ALX₃ compounds (AL2X, see also below for a discussion of B3LYP results). For these sets, the DFT-NL method reduces the error, compared to DFT-D3, by about a factor of 2.

These results are very promising. However, one should keep in mind that the NL term (and similarly the D3 correction in the BJ form) is attractive for short interatomic distances, which might lead to double-counting of electron correlation. The basic properties section of GMTKN30 consisting of, e.g., atomization energies, ionization potentials, electron/proton affinities, and barrier heights of small molecules can be used to test this. As can be seen in Figure 1d, there are indications for such double-counting effects which are, however, small. Only for revPBE and B3PW91 we find slightly worse results for DFT-NL compared to plain DFT, and for B3LYP both corrections even improve upon DFT. We have tried to analyze this in more detail by, e.g., inspection of the results for the atomization energies (W408) where the dispersion energies of the atoms are small (or zero for D3) and the molecules are systematically stabilized (as also seen from shortened covalent bond lengths in ref 17).

While for some functionals the MAD for W408 increases by about 30% (rPW86PBE and BLYP) with the NL correction, B3LYP improves slightly (from 4.3 to 3.9 kcal/mol), and revPBE0 benefits a lot (decrease from 9.8 to 5.5 kcal/mol). Apparently, this behavior is rooted in the sign of systematic errors of the uncorrected functional, i.e., already overbinding functionals become worse with NL terms and vice versa. This is also the reason why barrier heights of small systems (BH76 set) are mostly worse (typically, the MAD increases by 20%) with dispersion corrections: due to self-interaction error, the values are already systematically too low with the tested plain functionals, and the more compact (dense) transition states are further stabilized by the dispersion corrections. Note that this does not hold for the peri-cyclic reaction barriers (BHPERI), which are mostly better with D3/NL corrections. We attribute this to much lower self-interaction errors compared to BH76, where mainly atoms or small molecular fragments are involved. Overall, we think that the mentioned double-counting effects are present but smaller than one might expect, and in no way a reason to dismiss the corrections for the vast majority of applications.

Table 1. MAD (kcal/mol) for the S22 and S66 Non-Covalent Interaction Benchmark Sets with Dispersion Corrected Functionals

	S22		S66	
	NL	D3	NL	D3
B3LYP	0.48	0.31	0.48	0.32
B3PW91	0.45	0.47	0.50	0.33
BLYP	0.40	0.25	0.40	0.21
rPW86PBE	0.37 ^a	0.30	0.45	0.35
revPBE	0.29	0.44	0.21	0.31
revPBE0	0.18	0.28	0.24	0.22

^a The deviation to the value of 0.31 kcal/mol given in ref 17 results from different basis sets employed (def2-QZVP/noCP vs aug-cc-pVTZ/CP), a different type of integration grid, a non-self-consistent treatment of the electron density, and the use of slightly different reference values which we take from the recent S22 revision.⁴²

Because B3LYP is still widely used and moreover representative for many short-range over-repulsive functionals, we discuss its performance for the individual sets of GMTKN30. The results are shown in Figure 2. From the six VV10-type functionals, B3LYP-NL performs best for the complete set, and the WTMAD of 3.2 kcal/mol can be considered as being good for a hybrid with a relatively low fraction of Fock exchange. The best available hybrids yield values of 2.2–2.5 kcal/mol for this property; the average WTMAD for 23 D3-corrected hybrid functionals is 3.3 kcal/mol.⁵ Especially for reaction energies and basic properties, B3LYP-NL performs well while still being average for NCIs. From Figure 2, we note that B3LYP benefits most from the correction for intermolecular dispersion (IDISP), bond separation reactions of saturated hydrocarbons (BSR36), Diels–Alder reactions (DARC), difficult cases for DFT (DC9), and isomerization reactions of large organic molecules (ISOL22). For these sets, B3LYP-NL also clearly outperforms B3LYP-D3, which indicates that electron-density related effects are crucial for these difficult reactions. This view is further supported by the good results for the challenging systems in the “mindless reactions” (MB08165) set. The reason for the good performance of B3LYP-NL seems to be its short-range over-repulsive behavior leading to less double-counting effects. However, as already mentioned, some of these still remain, as can be seen for water clusters (WATER27), reaction energies of small systems (G2RC), barrier heights (BH76), ionization potentials (G21IP), and electron affinities (G21EA). Other sets that are slightly worsened are O3ADD6 (systems with multireference character) or sets related to self-interaction error (SIE11 or BH76).

Finally, we want to present results for the very popular S22³⁰ benchmark set for NCI. It has recently been extended to a more representative collection of complexes by the Hobza group⁴¹ (S66), and we take it here as kind of a cross-check for the performance for NCIs.

The rows are ordered according to decreasing MAD for S22 with DFT-NL. As can be seen from Table 1, all methods perform very well, as evident from MADs < 0.5 kcal/mol. The MADs without corrections are larger by a factor of 5–10, which makes the plain functionals almost useless for the study of NCIs. The MADs with the NL correction for S22 and S66 are highly correlated with each other. For both sets, the revPBE-type functionals perform best. For S66, the other four functionals

perform similar but somewhat worse. With D3, the performance ordering is similar for the S66 set (except for BLYP, which is the best D3-functional), while no such relation is seen for S22. Overall, revPBE and revPBE0 are very good with both corrections for NCIs, but all other combinations can be recommended as well. These data further indicate that the choice of the fitting set for determining the value for the parameter b should have only a minor impact on the results. We also note in passing that the theoretically motivated choice for rPW86PBE (being more Hartree–Fock like²¹) as a GGA component in VV10 (and vdW-DF2) does not lead to markedly better accuracy in these benchmarks compared to the other GGAs.

CONCLUSIONS

In this work, we assessed the performance of various nonlocal VV10 dispersion energy functionals with the GMTKN30 thermochemical database. We augmented the (hybrid)GGA functionals BLYP, revPBE, B3LYP, B3PW91, and revPBE0 with the individually refitted nonlocal (NL) part of VV10. All VV10-type functionals perform very well for the complete GMTKN30 database and achieve an overall accuracy comparable to that of the popular atom pairwise DFT-D3 correction. The density-based dispersion correction yields smaller errors for some important chemical reaction energies. This high accuracy comes at no expense in the description of noncovalent interactions, as tested in detail for the very representative S66 set.

Possible electron correlation double-counting effects are discussed but overall found to be of minor importance. In agreement with ref 17, we find that the VV10 correction works best with functionals that give no significant binding for van der Waals complexes. It can be expected that slight readjustments or even a simple increase of the amount of Fock exchange in the hybrids (because GGA exchange functionals also mimic correlation effects) might remedy most of the mentioned double-counting effects.

Our study shows that B3LYP-NL provides the best accuracy for reaction energies, basic properties, and the complete GMTKN30 database compared to the other tested VV10-type functionals, while still being reasonable (but not optimum) for noncovalent interactions. In that form, it can be recommended as a general-purpose electronic structure method. However, these overall very exciting results suggest that even higher accuracy might be obtained by combining the NL correction to better, specially designed semilocal functionals. Work in this direction is in progress in our laboratory.

Currently, DFT-D3 still remains the method of choice for many applications because it combines high accuracy with computational efficiency, in particular for routine structure optimizations. It further facilitates analysis and understanding of the dispersion contributions, which seems to be more difficult at a nonlocal, density-based level. Nevertheless, the very good agreement between the two conceptually very different dispersion corrections is satisfying. We seem to have achieved a basic understanding of electron correlation effects that were previously missing in DFT. At the current stage of knowledge and technical implementation, we can recommend DFT-D3 structure optimizations/dynamics checked by DFT-NL single-point computations. Future research should also investigate how both corrections describe nonequilibrium regions of potential energy surfaces and in particular metal-containing systems for which DFT-NL seems to be preferable.

ASSOCIATED CONTENT

S Supporting Information. GMTKN30 benchmark database and non-covalent interaction energy benchmarks. This material is available free of charge via the Internet at <http://pubs.acs.org/>.

AUTHOR INFORMATION

Corresponding Author

*E-mail: grimme@thch.uni-bonn.de.

ACKNOWLEDGMENT

This work was supported by the Deutsche Forschungsgemeinschaft within the framework of the SFB 858 (“Synergetische Effekte in der Chemie—Von der Additivität zur Kooperativität”). W.H. thanks Dr. Lars Goerigk for helpful discussions concerning the GMTKN30 database. We are especially grateful to Dr. Christian Mück-Lichtenfeld for his technical support.

REFERENCES

- (1) Stone, A. J. *The Theory of Intermolecular Forces*, 1st ed.; Oxford University Press: New York, 1997; Chapter 4, pp 50–63.
- (2) Kaplan, I. G. *Intermolecular Interactions*, 1st ed.; J. Wiley & Sons: Chichester, England, 2006; Chapter 2, pp 44–50.
- (3) Grimme, S. *Comput. Mol. Sci.* **2011**, *1*, 211–218.
- (4) Goerigk, L.; Grimme, S. *J. Chem. Theory Comput.* **2010**, *6*, 107.
- (5) Goerigk, L.; Grimme, S. *Phys. Chem. Chem. Phys.* **2011**, *13*, 6670.
- (6) Goerigk, L.; Grimme, S. *J. Chem. Theory Comput.* **2011**, *7*, 291.
- (7) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. *J. Chem. Phys.* **2010**, *132*, 154104.
- (8) Zhao, Y.; Truhlar, D. G. *Acc. Chem. Res.* **2008**, *41*, 157–167.
- (9) von Lilienfeld, O. A.; Tavernelli, I.; Röhrlisberger, U.; Sebastiani, D. *Phys. Rev. Lett.* **2004**, *93*, 153004.
- (10) Becke, A. D.; Johnson, E. R. *J. Chem. Phys.* **2005**, *122*, 154101.
- (11) Tkatchenko, A.; Scheffler, M. *Phys. Rev. Lett.* **2009**, *102*, 073005.
- (12) Sato, T.; Nakai, H. *J. Chem. Phys.* **2009**, *131*, 224104.
- (13) Steinmann, S. N.; Corminboeuf, C. *J. Chem. Theory Comput.* **2010**, *6*, 1990–2001.
- (14) Dion, M.; Rydberg, H.; Schröder, E.; Langreth, D. C.; Lundqvist, B. I. *Phys. Rev. Lett.* **2004**, *92*, 246401.
- (15) Thonhauser, T.; Cooper, V. R.; Li, S.; Puzder, A.; Hyldgaard, P.; Langreth, D. C. *Phys. Rev. B* **2007**, *76*, 125112.
- (16) Lee, K.; Murray, E. D.; Kong, L.; Lundqvist, B. I.; Langreth, D. C. *Phys. Rev. B* **2010**, *82*, 081101.
- (17) Vydrov, O. A.; Van Voorhis, T. *J. Chem. Phys.* **2010**, *133*, 244103.
- (18) Hujo, W.; Grimme, S. *Phys. Chem. Chem. Phys.* **2011**, *13*, 13942.
- (19) Koide, A. *J. Phys. B* **1976**, *9*, 3173.
- (20) Grimme, S.; Ehrlich, S.; Goerigk, L. *J. Comput. Chem.* **2011**, *32*, 1456.
- (21) Murray, E. D.; Lee, K.; Langreth, D. C. *J. Chem. Theory Comput.* **2009**, *5*, 2754.
- (22) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865.
- (23) Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098.
- (24) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785.
- (25) Zhang, Y.; Yang, W. *Phys. Rev. Lett.* **1998**, *80*, 890.
- (26) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648.
- (27) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. *J. Phys. Chem.* **1994**, *98*, 11623.
- (28) Ernzerhof, M.; Scuseria, G. E. *J. Chem. Phys.* **1990**, *110*, 5029.
- (29) Adamo, C.; Barone, V. *J. Chem. Phys.* **1999**, *110*, 6158.

- (30) Jurecka, P.; Spöner, J.; Cerny, J.; Hobza, P. *Phys. Chem. Chem. Phys.* **2006**, *8*, 1985.
- (31) Perdew, J. P.; Wang, Y. *Phys. Rev. B* **1986**, *33*, 8800.
- (32) Vydrov, O. A.; Van Voorhis, T. *J. Chem. Phys.* **2009**, *130*, 104105.
- (33) Vydrov, O. A.; Van Voorhis, T. *J. Chem. Phys.* **2010**, *132*, 164113.
- (34) Vydrov, O. A.; Van Voorhis, T. *Phys. Rev. A* **2010**, *81*, 062708.
- (35) Kolb, B.; Thonhauser, T. *Phys. Rev. B* **2011**, *84*, 045116.
- (36) Johnson, E. R.; Becke, A. D. *J. Chem. Phys.* **2005**, *123*, 024101.
- (37) Johnson, E. R.; Becke, A. D. *J. Chem. Phys.* **2006**, *124*, 174104.
- (38) Neese, F.; Wennmohs, F.; Hansen, F.; Kossmann, S.; Klamt, A.; Sinnécker, S. *ORCA*, version 2.7, Universität Bonn: Bonn, Germany, 2009.
- (39) Neese, F. *J. Comput. Chem.* **2003**, *24*, 1740.
- (40) Grimme, S. *J. Comput. Chem.* **2006**, *27*, 1787–1799.
- (41) Řezáč, J.; Riley, K. E.; Hobza, P. *J. Chem. Theory Comput.* **2011**, *7*, 2427.
- (42) Takatani, T.; Hohenstein, E. G.; Malagoli, M.; Marshall, M. S.; Sherrill, C. D. *J. Chem. Phys.* **2010**, *132*, 144104.

A Multibox Splitting Scheme: Robust Approximation For *ab Initio* Molecular Dynamics

Stas M. Avdoshenko*

Institute for Materials Science and Max Bergmann Center of Biomaterials, Dresden University of Technology, 01062 Dresden, Germany

ABSTRACT: In this work, the multibox (M-box) simulation scheme is introduced, which can be considered as a generalization of the QM/MM scheme for multifragment (molecular) systems. This scheme exploits the natural locality of multifragment molecular-based systems by mapping the system into force-coupled block subspaces. Where defined in this way, the entire system can be fully modeled under a quantum mechanical force field. This allows the description of each subspace explicitly by means of a robust electronic structure theory without the requirement for large computational resources. An adequate block-to-block coupling by means of shared subsystem fragments is applied to preserve the long-distance structural correlation in the system during a molecular dynamic (MD) simulation. Since electronic structure descriptions play a central role in the formulation of several parametric models for charge or energy transport, we expect that this space–time correlated scheme can become a reliable computational tool for charge/energy transport/transfer applications. The efficiency of the method is demonstrated by performing statistical and time-resolved analysis using both the multifragment box and full *ab initio* approaches. We illustrate the method using as examples the melting process of a one-dimensional benzene chain (weak interaction situation) and NVE dynamics for the C_nH_n polymeric chain (strong interaction situation). We also have extended the threshold of applicability of our model, demonstrating how it can cope with MD simulation with more complex systems and processes.

1. INTRODUCTION

The importance of molecular dynamics (MDs) in revealing the physics behind molecular and nanoscale phenomena has been recognized since the 1960s¹ when biophysicists performed the first X-ray experiments on enzymes. Nowadays, it is hard to imagine a modern X-ray data analysis of protein crystals without MD tools.^{2,3} Now, in the nanoscience era, MD simulations play a crucial role in current state-of-the-art nanoscale energy/charge transport/transfer models.^{4–6} The general idea in these studies is to combine a *fermionic* degree of freedom with time-resolved atomic motions (which are *bosonic*) to predict transport characteristics in molecular-based systems. As previously demonstrated,^{4–6} not only is the statistical average in the ensemble of interest important for such dynamic analysis but also the specific behaviors occurring on a short time scale domain (usually ~ 10 fs) are extremely relevant for a successful theoretical description of the electronic systems. On top of that, it is already known that for small systems (up to a few hundred atoms), Born–Oppenheimer (BO) *ab initio* MD techniques are enough to cope with the evolution of the electronic wave function along the dynamic trajectory of the ions and roughly estimate possible paths of the evolution of molecular orbital ensembles. However, the size of typical molecular-based devices such as organic light emitting diodes (OLED), organic field-effect transistors (OFET), or even transport/transfer effects taking place in certain biosystems generally includes many atoms (usually exceeding 1000 atoms). Although development of computational hardware broadens the ranges of the objects which can be studied by *ab initio* techniques with the use of super-computer resources,^{7,8} such techniques are still hardly applicable for describing structural dynamics of extended systems in a routine way. It means that *ab initio* models are already hardly applicable for describing structural dynamics of such extended systems.

Therefore, the implementation of an effective electronic structure description capable of dealing with such intricate ingredients is widely required.

A good electronic structure description is strongly linked to an accurate structural prediction and vice versa. Any tiny error in the first stage can lead to absolutely misleading parameters at the end, interfering strongly with the final transport or transfer characteristics. For instance, depending on the level of theory, a deviation of $\sim 1\%$ in the benzene molecule geometry can result in an overestimation of 0.5–1.0 eV of its molecular orbital gap size. At the same time, highly accurate approaches such as MP2, MP4, multiconfiguration models, or even hybrid functionals are impractical to employ in large systems with more than 1000 atoms, if the system is a bit complex (transition metals, conjugated systems). A reasonable solution for systems with a large amount of atoms would be the use of classical force field methods. Still, such classical techniques hold disadvantages since they are either not valid or fail for phase space integration on the short time domain.^{9,13,14} Also, such rough approximations following canonical parametrization types can estimate the wrong geometrical parameters for certain physical aspects such as molecular contacts or conformations.⁹

An alternative solution to improving the results could be achieved by implementing *ab initio* refinements on the snapshots of the classical trajectories,¹⁰ although some critical comments about this approach have been published.^{11,12} Problems will arise strongly for not so complicated systems such as the DNA backbone group PO_4^{-3} , noncanonical atom types (fluorine, boron, and again metals), and heterocycle compounds.⁹

Received: April 20, 2011

Published: October 24, 2011

Reasonable results have been achieved within semiempirical models (AM1, PM3, PM6)¹⁵ or the DFTB¹⁶ approximations. For structures close to their equilibrium, these methods are sufficiently accurate in obtaining the geometric and electronic structures. However, these models have limitations when MD simulations are conducted following the parametrization schemes of the respective methods. In general, the parameters are tuned on the basis of *ab initio* results; i.e., they are taken from reference equilibrium points located in potential energy surfaces which contain a limited amount of canonical interaction types.^{18,19} This compromises their transferability; hence nonequilibrium states or more complex molecular structures are not necessarily well described by means of semiempirical schemes.^{20,21} As a consequence, a verification of the parametrization set obtained for a particular molecular system is essential and often time-consuming.

The solution to overcome all of these problematic issues that we list here seems to be a return to the foundations of *first principle* methods.

The development of linear scaling methods has a long history. Hartree–Fock (HF), the post-HF (MP2, MP4, etc.), and DFT-based models have been widely used in the scientific community. On the basis of Rokhlin's ideas aimed at electronic structure problems,¹⁷ Scuseria and Kudin have implemented a direct space fast-multipole technique which was adapted for electronic structure problems, solving the Coulomb problem linearly for rather large system sizes.^{17,22,23} In parallel, Hutter et al. have developed a similar but more elaborate method of solving electronic structure problems by means of the Gaussian and Plane-Wave (GPW) technique.^{24,25} In their model, a dual density treatment of the Kohn–Sham model is applied to map the large-scale Coulomb problem onto Fourier-space through a plane-waves representation. At the same time, the one-electron terms can be efficiently treated using a local Gaussian basis. Another important contribution for the development of linear theories was performed by Gao in the X-pol^{26–28} model. X-pol forms the foundation for the Kitaura–Morokuma–Fedorov energy decomposition ideas summarized later in the fragment molecular orbitals (FMO) scheme.^{29–31} These models are highly suitable for a precise treatment of the local electronic structure in large-scale and macromolecular systems. The model takes advantage of splitting a multi-molecular system into coupled monomolecular quantum subsystems, which are coupled in a pure quantum manner. The model we are going to describe here has its origin in a system fragmentation detailed below. Importantly, in our model, each fragment acts as a separate QM/MM problem, and then a summation over crossed coupled subsystems is performed to obtain an approximation of the total system. In other words, the gradient of the whole system is derived from the multiforce treatment in terms of multi-QM/MM parts coupled due to shared fragments between different QM/MM problems, which closely approximates the gradients evaluated in the full *ab initio* description.

2. METHODOLOGY

2.1. Philosophy of Multibox Model: QM/MM. The Quantum Mechanics/Molecular Mechanics (QM/MM)^{32–35} approach is an example of a hybrid method that is currently one of the most popular models for the theoretical treatment of various physical phenomena in complex environments. In this method, the quantum region where all of the main chemical properties and reactions take place is treated quantum mechanically, while classical or semiclassical molecular mechanics are employed to model the

effect of the external potentials due to the surroundings. An example of a hybrid Hamiltonian for QM/MM problems is given by

$$\hat{H} = \hat{H}_{\text{QM}} + \hat{H}_{\text{MM}} + \hat{H}_{\text{QM/MM}} \quad (1)$$

where \hat{H}_{QM} deals with the quantum region, \hat{H}_{MM} gives the energy of the molecular mechanics contribution, and $\hat{H}_{\text{QM-MM}}$ represents the interaction between the QM and MM parts. In more detail, the term \hat{H}_{QM} can be a quantum operator acting on the space of N_{bs} local atomic basis functions, where the level of sophistication of the adopted theory is dictated by the size of this basis set. The computational cost required for modeling this term can be $O(N_{\text{bs}}^n)$ with $n = 2–8$, depending on the level of theory. The \hat{H}_{MM} part consists of classical interactions located in the surroundings, and its computational complexity can reach $O(N_{\text{a}}^2)$ where N_{a} is the number of atoms. In some cases, more efficient approximations for the potentials and cutoff functions are capable of reducing this numerical complexity to $N_{\text{a}} \ln(N_{\text{a}})$. The graphical representation of the typical QM/MM model is shown in Figure 1A.

One of the advantages of QM/MM is the possibility to describe the environmental effects in a reliable way since an explicit contribution due to the presence of solvents, for instance, can be included in the theory easily. Nevertheless, the description of the QM region, which carries most of the interesting physics, can be significantly compromised since its efficiency still scales with $O(N_{\text{bs}}^n)$. However, for complex systems containing about 1000 atoms in the QM region, this model becomes impractical at more sophisticated levels of theory. In other words, the size of the QM region restricts the applicability of this approach to the same extent as pure assumed *ab initio* techniques.

However, we can easily split such big system in parts following Gao's fragment picture and construct the QM/MM problem for each subsystem and then find a way to merge these subsystems to reconstruct a more reliable approximation for whole original system space. In this fashion, all long distance effects will be defined locally but, at the end, are collected over the entire system.

2.2. Multibox Origin: X-pol Method and Fast Multipole Model. The multibox (M-box) scheme introduced here provides a way to overcome some of the restrictions imposed while dealing with large QM systems. The main idea is to apply fragment orbital concepts mixed with the QM/MM formalism^{26–28} and make these fragments coupled in the sense of forces coupling instantly between shared elements of different fragments. For simplicity's sake, we illustrate this approach in linear homomolecular systems; nevertheless, the generalization of the ideas presented here to heteromolecular or topologically different species is straightforward.

Following the ideas of Gao,^{26–28} it is possible to exploit the natural locality of molecule-based structures with strong localized electronic structures by treating each molecule as a separate QM system. This implementation already reduces the computational cost of the problem to $O(M^*(M_{\text{bs}})^n)$, where M is the number of molecules in the system and M_{bs} is the number of basis functions for a single molecule. However, fragmenting the system into pieces can become a real problem when dynamic aspects need to be treated. First, the repartition of the system into weakly interacting QM regions removes the dependency of the electronic structure on the environment and nonlocal effects. Such a splitting scheme creates M independent trajectories in the phase space with only weak QM interaction considered. This decoupling procedure is invalid when charges/spins or vibrations are correlated over a long range in space (on a larger separation than

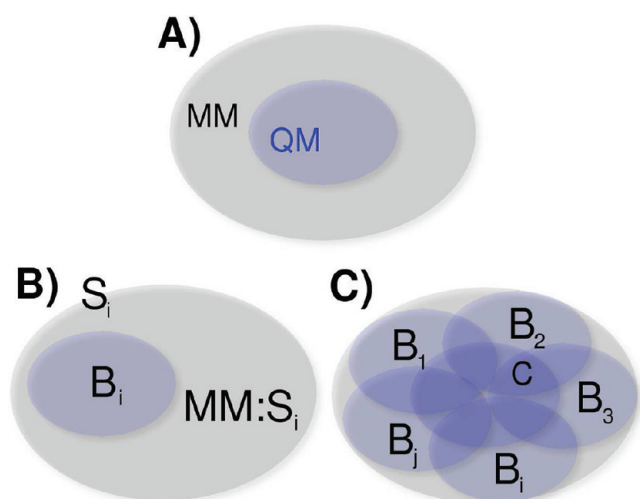


Figure 1. (A) Graphical representation of QM/MM model concepts. The graphical representation of the M-box model. (B) Definition of the subspace S_i containing the quantum block B_i , using QM/MM charge effects from neighbors $MM:S_i$. (C) The whole task mapping, where the system is covered by active QM space. The border (C) is a crossing of the different QM regions.

the fragment). In this sense, one needs to preserve an effective short-range correlation description while at the same time including long-range effects without compromising the computational efficiency.

2.3. Multibox. In the M-box model, the dependence of the molecular blocks on the environment is constructed in two ways.

First, following the philosophy of the tight binding (TB) approximation, a better description of the short-range interactions can be obtained by maintaining at least nearest neighbors contribution on each QM subsystem. Therefore, the system is not simply split into pure isolated single-molecular blocks but into effective blocks, which contain the information of the molecule “surroundings” in the full quantum sense. In this way, there is no need to treat the system within the FMO approach since neighboring effects can be naturally included as sketched in Figure 1. However, as in the FMO approach, the use of a broad spectra of theoretical methods for this effective fragment is possible.

At the same time, for each block, the influence of long-range Coulomb interaction can be implemented via the standard QM/MM method, e.g., by means of a classical electronic coupling (point charge scheme in the simplest case). It is important then to compare the alternative method presented here with the pure QM/MM description on the same QM scale. In parallel, one can show explicitly how the construction of the subsystems is done within each methodology. It is trivial that, in some cases, simple electrostatic corrections due to the influence of certain solvents are already sufficient for a proper description of the systems, and quantum contributions are not that important. Additionally, as was found for the fast multipole model (FMM)—which is rather similar to QM/MM ideas—the solution for the electronic structure part can become more accurate. It follows from the fact that, for many sophisticated and high-level theoretical chemistry models which are based on the Cauchy-Schwarz inequality, Coulomb interaction screening cuts are defined parametrically for long-range contributions. On the other hand, in the FMM/QM/MM-like picture, all nonlocal contributions are treated locally and exactly with the aid of the multipoles field.

Here, we start now describing the new splitting scheme, which is centered on placing more than one molecule in each block. The number of computational operations increases to $O(M_{bl}(N_{bl} * M_{bs})^n)$, where M_{bl} is the number of blocks and N_{bl} is the number of molecules in one block. Hence, the method scales linearly with the number of blocks, and it is only limited by the level of theoretical calculations performed on each local block, for instance, HF|MP2 or more sophisticated approaches. As the maximum number of blocks is equal to the number of molecules, one can state that the method still scales linearly with M . Furthermore, we assume that the molecules from other blocks are coupled to the reference block by the construction of QM/MM interactions established for each subsystem. In addition, it is also possible to simply consider that each block feels a classical Coulomb response caused the MM parts of blocks’ representation.

Following this setup, it is ensured that each block interacts with the whole system at least at the level of electrostatic coupling. In other words, a set of M complete but independent QM/MM systems, S_i , is constructed from the original system, S . At this point, we have to emphasize that the model described here is closely related to Gao’s ideas, with the extra ingredient that, instead of decomposing the energy in the space of the molecular fragments, we have constructed topologically identical multisets which are linked through the forces in the shared parts of the fragment.

2.4. Multibox Forces and Hamiltonian. However, the time evolution of the independent blocks in the M-box approach remains uncorrelated with phase space. The pure inclusion of long-range Coulomb interaction by means of classical charge effects—all blocks mutually affect each other—is not enough to guarantee the space–time correlation. Additional ingredients must be implemented to solve this issue; in particular, one can employ broad spectra schemes for possible dispersion effects (classical VdW, partial classical fields to describe long-range interaction, classical charge effects, etc.).

For this purpose, we introduce a number of molecules with special *interconnect* properties. Instead of creating a system composed of isolated blocks, one assumes that each block shares a part of its own atomic structure (via an interconnector) with its neighbors. Formally, one can say that these molecules act on different subsystems within the same time scale, thus controlling the evolution of the whole system. This leads to the definition of a novel *coupling model* which can be redefined depending on the problem of interest, or modeling task (transport/transfer). In the simplest case, the coupling model can just be built up from a linear combination of the *ab initio* forces, \vec{F}_I , acting on the molecule I located in different blocks, B_j , with J specifying the central molecule in the block. Thus, the coupling model can be read as

$$\vec{\mathcal{F}}_I^S = \begin{cases} \vec{F}_I & \text{if } B_I \in \mathbf{B} \\ \sum_{I \in B_j} \vec{F}_I^{B_j} & \text{otherwise} \end{cases} \quad (2)$$

with $\vec{\mathcal{F}}_I^S$ being the final force on the molecule and \mathbf{B} labeling the complete set of blocks. The fact that the molecules belong to different QM blocks indeed increases the complexity of the method. This compensates for the force linearity in M . In analogy to QM/MM methods, a graphical representation of the method can be seen in Figure 1.

A natural question which could be raised at this point is whether the validity of the Hamiltonian principles is maintained (e.g., conservation of energy). On the basis of the ideas described above,

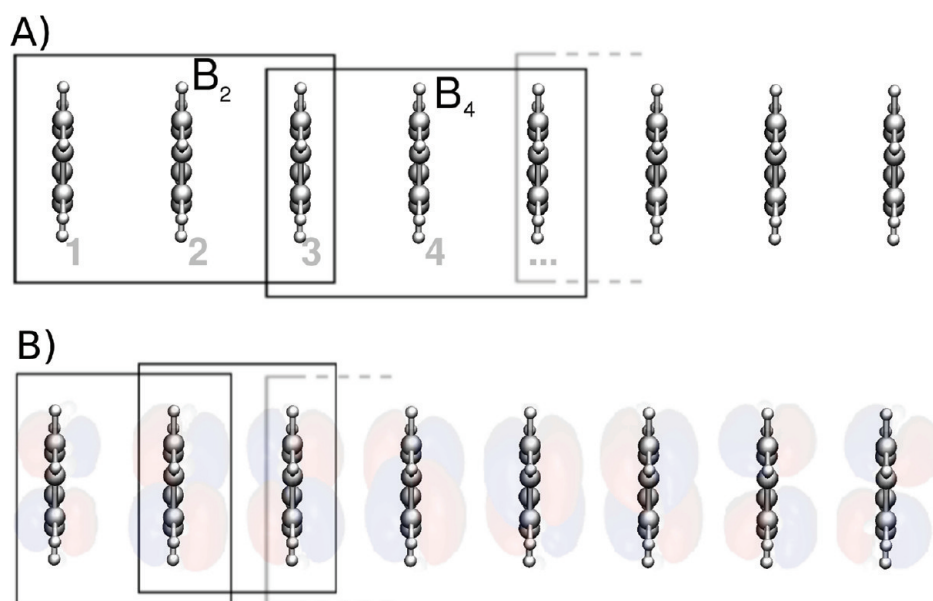


Figure 2. (A) The 1D benzene chain with highlighted QM regions according to the splitting applied in the simulations. Also shown are the coupling sites ($3, 5 \in B$) between B_2 and B_4 and between B_4 and B_6 in the frame of the M-box model. (B) The 1D benzene chain splitting with natural system symmetry preservation. Block-HOMO orbitals for one of the initial trajectory point are shown.

one can see that the M-box description operates with partial forces which are naturally close to a constrained Hamiltonian scheme.⁴⁶ In fact, this is the idea behind the coupling method: to *glue* the systems by these common parts in terms of constrained forces and try to preserve this mapping back to the original space in the integration scheme.

That is why, as we are interested in the evolution of system forces for only N particles, a reasonable choice for constrained forces can be determined by chemical intuition. An assumption could be that the potential energy of the whole system (original on $2N$ -space, or N particles) can be represented as a sum of other blocks (b). Then, one can derive explicitly the partial forces ($F_i = \sum_{k \in b} \partial / \partial q_i U_k(q)$ for $i \in N$), and the partial moments $p_i = \int F_i dt$, where $U_k(q)$ is the potential energy with respect to the subsystem $k \in b$. From this, the definition of the system forces becomes a trivial problem.

3. RESULTS

3.1. Multibox vs. Full BOMD: Weak Interaction Limit. In the following, the quality of this approach will be demonstrated by a direct comparison of Mbox results with full *ab initio* BOMD results. To illustrate this point, we adopted a unidimensional (1D) periodic chain composed of benzene rings in order to avoid additional complexity in this first analysis. At the same time, a 1D system is not a totally arbitrary choice since in many organic systems (crystals), structural characteristics often have a principal axis and well-known properties such as anisotropy according to these axes. Hence, many 3D structures can be deconvoluted into a 1D stack.

Figure 2 shows the system of interest with eight molecules in total. It also depicts two possible ways in which the system can be split. In panel A, the system apparently reveals symmetry breaking (ring-to-ring distance is ± 0.005 Å) which does not occur in the full (eight molecules) system optimization. In panel B, the system is fully symmetric (ring-to-ring distance is almost the

same as in the full system optimization). Since intermolecular forces are very weak and only local impacts are essential, it is not surprising that in both cases a good agreement is achieved with results for the full system QM optimization.

We now discuss the molecular dynamic simulation results based on system A, where each block contains three molecules, and there is only one molecule in the coupling region. In the coupling region, the forces for each block are read as $\vec{F}_3 = \vec{F}_3^{B_2} + \vec{F}_3^{B_4}$ in the case of the molecule labeled 3. The initial system guess was made using a ~ 3.8 Å translation vector of the original equilibrium C_6H_6 structure, with preoptimization until a system gradient of 10^{-5} au/Å was obtained. In this case, the system is small enough to fully treat it using *ab initio* BOMD for comparison with the M-box result.

The benchmarks were considered using the GPW-based³⁶ CP2K code,³⁷ which allows a mapping of the system onto a different force environment. Also in this code the QM/MM interface is implemented by its developers.^{38,39} All MD simulations have been performed in the microcanonical ensemble (NVE) in order to avoid random velocity renormalization and friction forces from thermostats. A time step of 0.5 fs and an initial temperature of 300 K was applied during all of the simulations. For each subsystem, the QM/MM problem was solved where the properties of the QM part were defined at the DFT|PADE|TZV2P⁴¹ level of theory with a density cutoff of 280 Ry in the field creating by an external MM electrostatic potential.

For the classical MM part, the AMBER-99^{42–44} force field was used for the aromatic benzene atoms. We emphasize beforehand the fact that in the current study the classical dispersions do not influence the evaluated forces. This can be done easily by changing the coupling properties within each block and between blocks. In accordance to the M-box scheme, each subsystem acts under QM and MM force fields, so this gives rise to the generalized scheme of coupling forces as a weighted superposition of classical and quantum forces. With these computational options, the energetics and geometrical characteristics obtained from the full

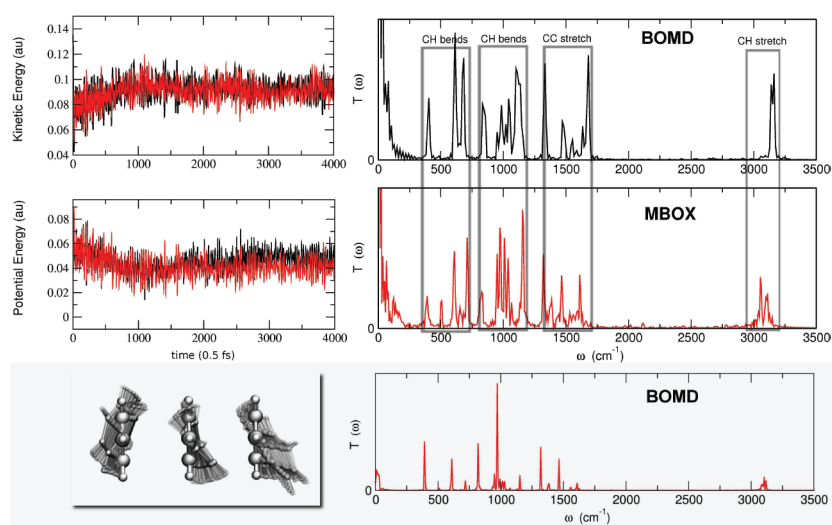


Figure 3. The evolution of the kinetic and potential energies (absolute value for both runs with respect to the start point) during BOMD and M-box run over 2.0 ps (left panel) and VDOS spectra function of the kinetic energy for 2 ps production runs for both models, where characteristic band regions are highlighted (right panel). The 2 ps region has been chosen because on this scale all other molecules of the chain are still in close vdW contact. (Lowest gray panel) VDOS based on time 2 ps NVT evolution for a three-molecule system (single block in the M-box splinting), on the left-hand real space geometry changes during the first 1 ps NVE run with 20 fs sampling.

ab initio BOMD simulation are compared to M-box results. Figure 3 presents the short time evolution of the kinetic and potential energies. The data show that the deviation for both methods covered the same energetic range with a good correlation of the on-time behavior. The initial structure guess at 0 K and the random velocities distribution at 300 K distribution results in total kinetic energy varying around an expectable average value of $T \sim 150$ K.

The collective behavior and the significance of the *subsystems coupling* cannot be easily demonstrated in the real time domain. Therefore, the Fourier spectra $[T(\omega)]$ of the kinetic energy in both simulation runs based on 2 ps trajectories (right panels in Figure 3) are analyzed. The spectral function (VDOS) can be compared with the well-known vibration spectrum for the benzene molecule, which has only four prominent IR bands in the gas phase resulting from the dipole selection rules.⁴⁵ At 1500 cm^{-1} and $3100\text{--}3000\text{ cm}^{-1}$, the stretching modes associated with the aromatic carbon–carbon bonds and the aromatic CH appear, respectively. The bending motions involving carbon–hydrogen bonds appear at 1000 cm^{-1} for the *in-plane* bends and at 675 cm^{-1} for the *out-of-plane* bends. Both models are able to reproduce all of the mentioned bands. As one can see, M-box closely reproduces the spectral function obtained by the full BOMD description. It is remarkable that the low energy profiles up to 500 cm^{-1} for both methods are mostly identical to each, illustrating the close character of collective motions predicted. An additional observable band splitting can be associated with collective effects. It is closely related to the differences between the gas and liquid phase experimental IR spectra, where for the liquids, a much richer IR spectrum should be detected.

The impact of the subsystem coupling can be illustrated by analysis of the behavior of an independent subsystem behavior on the same time scale. Figure 3 (gray panel) shows the spectral function of a single subsystem which was obtained on the same time domain of 2 ps. Obviously, this spectral function differs dramatically with respect to the previous cases because of the lack of additional dispersion from the molecule–molecule interplay.

Again, high and low energy bands can be seen in this VDOS spectrum, but additional band splitting (band intercombination) is small and the phonon region (up to 500 cm^{-1}) becomes extremely thin.

In order to exclude the effects of random initial guesses in the frame of the NVE ensemble, test runs were analyzed for several independent runs. In these analyses, all observables (energy mean values, local dipole moment, etc.) were nicely reproduced. The real time geometrical changes averaged over 10 independent BOMD and M-box runs are considered. Each trajectory was started from the same structural initial guess but with different initial velocities. The resulting geometrical trajectories for both techniques are presented in Figure 4 with 50 fs sampling for the graphical presentation of a 2 ps trajectory. Due to the dissociative nature of the melting process in these tasks, a detailed statistical analysis (principle component analysis, for instance) is difficult to perform because of a large contribution of irreversible types of atomic motions. Nevertheless, from Figure 4, it is obvious that those average trajectories are almost identical, and one can reasonably assume a close space–time correspondence between the M-box model and the full BOMD models. Most important is the fact that in all of the above runs, the splitting of the whole system (A–symmetry break, B–symmetrical task) leads to an artificial symmetry breaking with respect to the proposed force-coupled QM regions. While the central atoms are in a symmetric environment as QM describes all of the neighbors, the coupled atoms are not. In each of our blocks, one of their nearest neighbors is part of the QM region, while the others interact via the QM/MM coupling, which is an electrostatic interaction only in this report. Being aware of this fact, the overall agreement with the QM results indicates that this different treatment of the molecules has no significant effect on the description method of the whole system. In other words, it means that the results are almost the same, independent of the splitting scheme (see Figure 2). Nevertheless, for more complex systems or to prevent a too long extension of this problem, a different definition of the quantum blocks should be implemented according to the system symmetry, by assuming that the environment is treated in the same way for

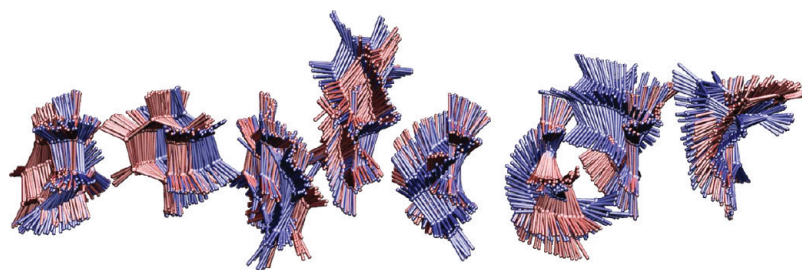


Figure 4. Superimposed real space changes for full BOMD (red) and M-box (blue) runs in a 2 ps trajectory window with 20 fs sampling (average over 10 independent runs).

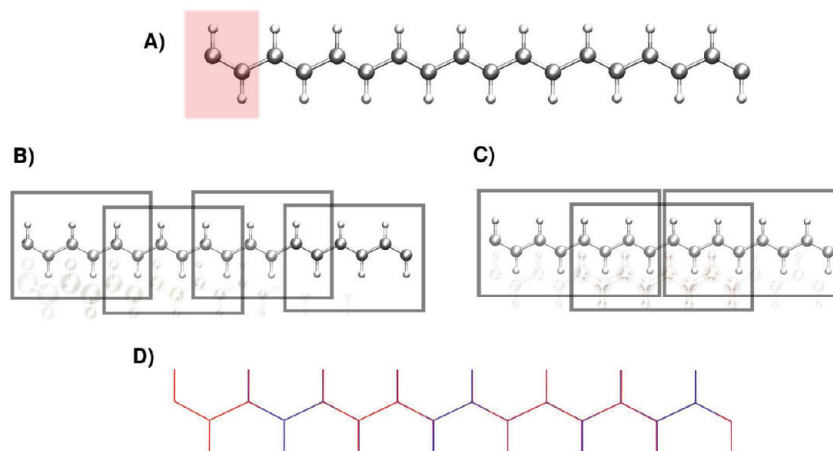


Figure 5. (A) The original systems of interest. Unit cell for the *trans*-(CH)_x chain, the red rectangle highlights two different origins ($x = 18$ and $x = 16$) for whole systems where Γ -point solutions are remarkably different. (B–C) The two chosen ways of system mapping, see text for details, depending on the origin, $x = 6$ (benzene like) and $x = 8$ in the sub-block. (D) The merging optimized structures obtained by using the M-box (red) scheme and full QM method (blue) for the (C) $-x = 8$ and system (A) $x = 16$ (see text for details), where a structure comparable in the sense of G -point results runs. The C–C bonds length are split at 1.39 and 1.47 Å, as it is easier for (A) $-x = 16$ and (C) $-x = 6$.

each molecule (or parts) of the system. This could be achieved, e.g., by defining the blocks such that, for every molecule I , a block B_I exists.

3.2. Multibox vs BOMD: Strong-Interaction Limit. In order to address all of the possible interaction limits, we discuss an interesting set of systems where a strong or covalent interaction between the blocks is considered. According to our initial goals, it is worth mentioning that the M-box scheme is capable of treating systems where weak interactions take place, and so the method would not be expected to work well for the problem with strong interactions.

As an example, a simple polymeric $-\text{[CH]}_x-$ annulene was chosen. One can see that such a problem involves a particular complexity which is associated with π systems. It is clear that the QM/MM standard scheme may face problems while describing bonds at the edge of the fragment, even if their saturation by means of QM/MM is conveniently addressed. Extra complications can arise from Γ -point calculations. On top of that, the pathogenic density delocalization in the DFT picture itself also becomes a difficulty.

This issue has been analyzed in the same M-box spirit as our results presented before. The $-\text{[CH]}_x-$ systems are shown in Figure 5, highlighting the several ways of mappings where different variances with distinct sub-block sizes and intercoupling have been chosen. The subsystems of interest were translated in such ways as to avoid the QM/MM saturation of QM-block

terminal atoms. This means that the QM cell size for each block is exactly sufficient to allow the saturation of this QM block by the translation of this block. This way, we try to reduce or completely eliminate some nonphysical contributions from the correlation spectra function which are often driven by this saturation.

We keep allowing the system to move with respect to the forces defined in each QM region and for every block and with respect to the coupling model defined for the cross-sections illustrated in Figure 5B and C.

It should be emphasized that variations of the properties of the polyacetylene as a result of the fragment taken as a translation unit is the outcome of Γ -point calculations. If the whole k space is covered, both approaches should give identical results. However, complete Brillouin zone integration at each MD step is impractical because of the high demand it would make on computational resources. Therefore, the proper choice of the translation element is crucial to make Γ -point computations reliable.

It is clear that for a hypothetical $-\text{[CH]}_x-$ annulene depending on $x = 18$ or $x = 16$, the Γ point can be related to $k = 0$ and $k = \pi/a$ states. In the chemical language, it can be explained in terms of Huckel's $4n + 2$ rule.⁴⁸ In fact, an N -site Huckel matrix is related to a nonperiodic system, such as the one addressed here. It means that the aromatic condition given by $(\alpha - \epsilon)/\beta = -2\cos(2\pi n/N)$ is valid. For a simple tight-binding Hamiltonian and for $N = 2, 6, 10, 14, 18, \dots, 4n + 2$, no antibonded contribution is possible until the limit is reached where the k space is reduced

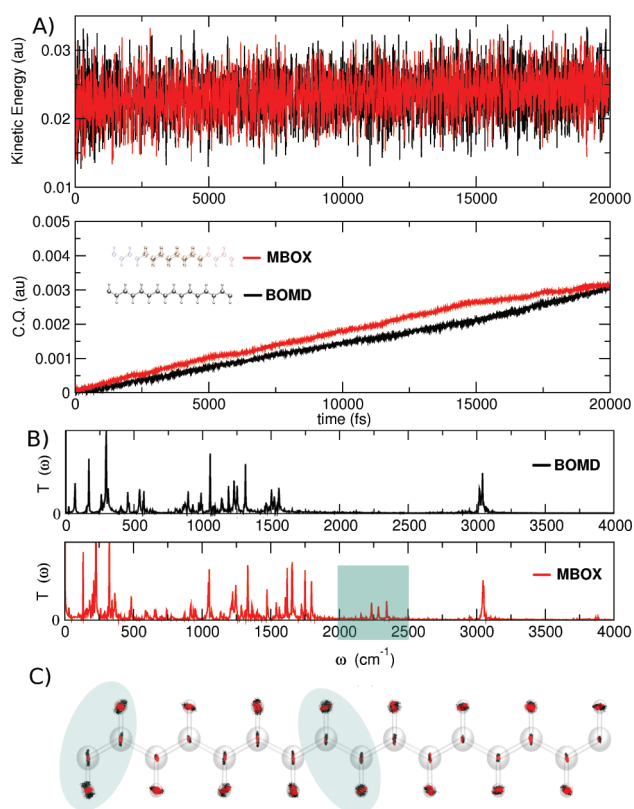


Figure 6. The deviation of the kinetic energy change during full BOMB and M-box run for 20 ps (A) and the VDOS spectra as a function of the kinetic energy for 2 ps production runs for both models, where characteristic band regions are highlighted (B). The real-time system behavior over a 10 ps trajectory with 50 fs sampling (C).

to a single point (Γ point). According to this, in the present cases, aromatic-like systems with sub-blocks containing $x = 6$ ([6]annulene) should be comparable with $x = 18$ ([18]annulene) and $x = 8$ ([8]annulene) with $x = 16$ ([16]annulene), e.g., nonaromatic systems.

The initial systems were optimized using the BFGS (Broyden–Fletcher–Goldfarb–Shanno) scheme for an equilibrium geometry search. For the search of equilibrium structures and MD study, the following splitting schemes into the blocks were applied:

- system B in Figure 5, with $x = 6$ elements in each block and $x = 2$ elements in the coupling region between blocks—the full system has $x = 18$ elements
- system C in Figure 5, with $x = 8$ elements for each block and $x = 4$ elements in the coupling region between blocks, which is logically comparable with whole structure with $x = 16$ elements structure

From our numerical observation, both choices have shown almost identically accurate final structures at 10^{-5} au/Å. It is remarkable that an intuitively rough approximation such as the one adopted in B was already enough to get the final equilibrium geometry close to the one predicted by the full QM theory. These runs have shown a good correlation for both splitting schemes, such as C–C bonds in chains being 1.37/1.49 Å (for the nonaromatic case) and 1.43 Å (for the aromatic case) after reaching the assumed tolerance factors. We will see later that such a split approximation as in B is not really applicable for the phase sampling, and case C seems to be more reliable in a statistical sense.

The MD runs for the strongly bounded regime have been integrated for a long time of ~ 20 ps. This has allowed us to make statistical and local analysis in a more precise manner. We will focus on case C with $x = 8$ elements in each fragment and $x = 4$ elements in the coupling region between blocks, and the system used for comparison is $x = 16$ in Figure 5.

Figure 6 shows the long time evolution of the kinetic and potential energies of the system following the M-box and full QM-BOMB descriptions. As one can see, the time variations of the kinetic/potential (here, the potential energy is given by the module with respect to the $t = 0$ fs point for the whole and M-box tasks in such a way that the starting point is apparently different) energies are identical. However, statistics show that, in contrast to the *weakly interacting case*, additional frequencies interfere with the reference spectra function from the full BOMB problem. Moreover, the VDOS functions are different in the range of 1500–3000 cm^{-1} (Figure 6). However, all of these new features in the M-box model can be straightforwardly understood.

We now analyze the shift of the C–C stretching modes from 1500 cm^{-1} predicted by full the BOMB to the more experimentally relevant range⁴⁹ of 1700–1800 cm^{-1} predicted by the M-box approach. The origin of this shift lies in the fundamental failure of DFT in overestimating the electronic delocalization. For instance, even in the case of the nonaromatic [16]annulene (e.g., $x = 16$), overestimation of delocalization by DFT in the full BOMB results in a noticeable weakening of the double bonds (which leads to an aromatic-like frequency at 1500 cm^{-1}), making such MD sampling problematic. In fact, this error is due to the shrinking atomic space within blocks of smaller atoms under the single particle DFT construction. Eventually at the same quality of basis (DZ/TZ), the M-box model might be more accurate in a prediction for smaller systems, where DFT methods are more suitable even if the system is strongly localized. We performed an additional analysis for [6]annulene, and results were mostly identical, and $x = 18$ ([18]annulene) around 1500 cm^{-1} is mostly identical. However from the VDOS spectra for the B approximation, one can say that the spectra shapes are similar for the M-box, and for the full BOMB simulation the characteristic bands (C–C and C–H) can be regarded from both models.

The most pronounced differences in the results can be seen around 2100–2200 cm^{-1} . Here, we need to conclude that the M-box scheme does not work perfectly well. Apparently, the periodic boundary conditions which were chosen to saturate the blocks are able to generate additional effects on the C–H dispersion (in the M-box, formally, we have eight terminal hydrogens) in the periodic reflection. These modes were not so efficient in case of the full BOMB problem (there are only two C–H bonds on the edges). This proportionality with respect to the number of the actual terminal groups can be seen as well in example B, where the intensities around 2100 cm^{-1} bands are much higher than expected.

It has been shown above that the ideas underlying this very simple scheme have the ability to retain the simplistic integration, and the system retains its Hamiltonian properties. It is well-known that numerical and energy definition problems such as the ones in DFT calculations have effects on the conservation quality (QC), which provides a drift in the sum $QC = E_{\text{pot}} + E_{\text{kin}}$ (Figure 6).

From our numerical observation for the vdW system (our benzene chain), conservation quantity has a slight deviation in comparison to the values of ~ 0.00050 au/ps obtained for the M-box and for the full BOMB problem, which is only ~ 0.00045 au/ps,

independent of the splitting schemes. This is due to the interplay of forces being defined only within a molecular pair. It is locally defined and affects the whole prediction, contributing additively over all errors given by BSSE, SCF accuracy, etc. at our level of theory. All of these errors in vdW complexes vary slightly with system size because of obvious reasons (one can think about this fact in terms of energy decomposition analysis, where it is clear that the contribution to BSSE given by four molecular centers, for example, is an irrelevant value). Hence, for our benzene system, the QC drift is simply defined in the benzene–benzene pair, and it is precisely the same for both BOMD and M-box.

The situation around conjugated chains develops differently, but of course it is driven by the same ideas as before. This is partially due to the impact of the bigger SCF wave function uncertainty predicted in larger systems in contrast to smaller ones, with the condition that close tolerances for the SCF procedure are used. In other words, it is simpler to obtain a more accurate approximated wave function for a small system. Because of this fact, for a comparable system size such as $x = 8$ ([8]annulene) blocks and its analog $x = 16$ ([16]annulene), they have QC deviation on the same energetic scale. However, in the short sub-block case—system $x = 6$ blocks and $x = 18$ —one can even have a smaller drift in QC values. This confirms that the SCF problems are scalable for the conjugated system (due to the band gap size problem, for example).

3.3. Multibox and Boxes Crossing Processes. An interesting investigation would be to verify what will happen if an event of particle exchange between boxes takes place. In other words, how much time-dependent particle exchange between blocks and along interconnect regions (B) is needed? Can this be described using our M-box scheme? A conceptually close approach, with a real-space splitting (with respect to the level of theory as well) for on-flight QM/MM border crossing processes has been discussed already by Buló et al.⁵² and earlier by Truhlar and Heyden.⁵³ For a system with strong interactions, such a crossing effect has to be very rare. Besides, these processes of interest should have a relatively large time scale where *ab initio* knowledge hardly can obtain information for the whole time interval in general. In this sense, only weakly interacting systems would need this knowledge, e.g., liquids, gases, etc. Systems with “weak interaction” can undergo large spatial rearrangements on a short time domain. A combined model can be represented as the following: a protein chain coupled block-by-block. In this structure, conformational transformation is not possible, and it holds a stable conformation for all of the simulation time. An example is closely related to the C_nH_n chain we have described before. Also, we can have the surrounding solvents shell some light and mobile molecules, which due to weak of interactions between them can cross the fragment boxes with a high probability. Here, we can construct a simple example like a light gas, which will expand in view of the NVE ensemble.

For the sake of simplicity we consider the ideal system dynamics of adiabatic gas expansion. According to the example examined below, we stress that one has to upgrade our level of theory to one which is grid-independent. It could still be a DFT model, but we need to isolate the plane wave part or make a careful recount in the renormalization. For obvious reasons, if the total energy of the system is a grid function such as in the PW part of a DFT model, one cannot avoid step-like changes in the total energy due to different grids for different amounts of particles under consideration. Hence, we use the AM1¹⁵ level of theory, which is also implemented in the CP2k code.³⁷ An ideal gas composed of Xe

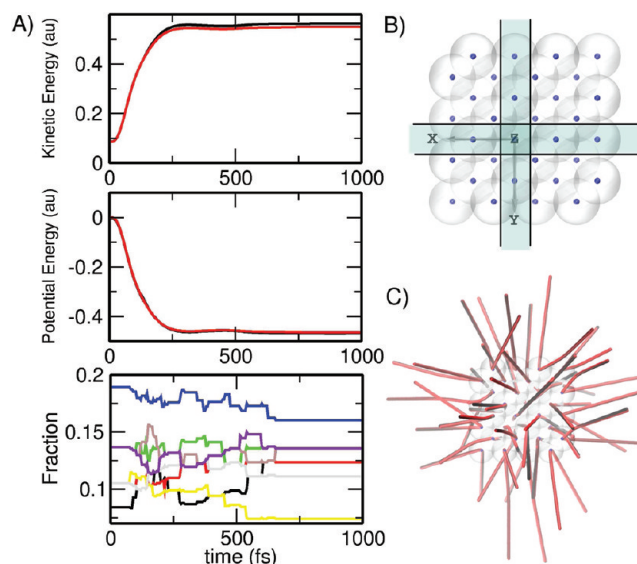


Figure 7. (A) Comparing the deviation of the kinetic and potential energies during full BOMD (AM1 level of theory) and M-box for the Xe-cluster tasks (B). Task setup with the highlighted B region in the real space and the box organization in the XY projection. (C) Real space Xe-cluster evolution for full BOMD (black) and multibox (red) runs in a 2 ps trajectory window with 10 fs of sampling.

atoms was assumed as the trial system, and its original geometry as shown in Figure 7 is taken from the diamond supercell with the Xe–Xe distance equal to 1.7 Å. Eight boxes have been constructed due to geometrical reasons. The system geometry is symmetric with respect to the (0,0,0) point. Cube 1, for example, is given by $(x > -1, y > -1, z > -1)$. Cube 2 is $(x < 1, y > -1, z > -1)$. This problem is defined as it automatically forms a 2.0 Å coupling region and B region, as can be seen in the 2D (XY) projection.

Our action script is next; every 5 fs, we checked which block each particle was supposed to be in with respect to the described splitting scheme. As occurred in previous cases, the blocks are coupled by the gluing (force merging), which is also dynamically mapped onto the full system size, which means that the temperature regime—or kinetic energy—is preserved by the limited number of atoms. However, within this construction (box crossing description), total potential energy given as an additive sum is not correct anymore and cannot be considered as an energetic criterion. For homogeneous systems, this inconsistency can be easily avoided by a renormalization of all of the parts which contribute to the total energy by using a weight of active parts normalized to practical numbers, e.g. $U_{\text{total}} = \sum_b w_b U_b$. One can estimate that if the system is equilibrated or optimized for a certain box size, the interesting event—box crossing—can take a long time. We accelerate the task by making this event highly probable on a short simulation time scale of 1 ps. In order to achieve this, we made the starting point cluster highly stressed, and then the system followed the NVE expansion. Our blank theoretical experiment involved the same initial geometric guess following the same AM1 theoretical model. The result of this construction is presented in Figure 7A (lower panel), where the evolution of kinetic and potential energies can be seen. As can be seen, the compressed gas with the high potential energy at the starting point is about to expand. The potential energy decay is accompanied by the expected growth of temperature/kinetic energy (similar time dependencies can be simply derived from

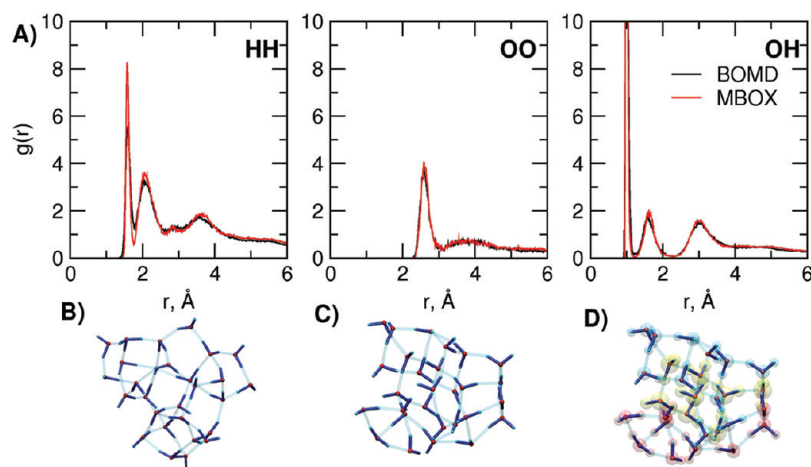


Figure 8. (A) Site–site distribution functions for water molecules (H–H, O–O, and O–H couples), reconstructed from the MD simulation with full *ab initio* and M-box models. (B) Water cluster geometrical characteristic after 4 ps of MD simulation in the case of full *ab initio*. (C) Water geometrical characteristics after 4 ps of NVT-MD simulation for the M-box model. (D) M-box splitting scheme (fragments 1, red; 2, yellow; 3, blue transparent), propagation done with the force vector defined as $\vec{F}_{\text{tot}} = \vec{F}^{1-2} + \vec{F}^{2-3} + \vec{F}^{1-3}$.

Lennard-Jones or Morses’s potentials). Obviously, the pictures are mostly identical for the M-box model and full description cases. Also, the real time changes are identical, as illustrated in Figure 7C, where a snapshot representation of the trajectories is shown. The small overestimation in the saturation limit for the M-box tasks can originate in the random nature of the initial velocity guess. As shown in Figure 7 by dynamics of fractions ($N_n(\text{inblock})/\Sigma_n N_n$), our goal was achieved, and during the simulation, the number of particles in the boxes was changed. Moreover, the potential and kinetic energy time evolution results seem to reveal that the conservation criteria are preserved, at least for the present NVE integration.

Besides the time evolution of the specific systems that were discussed previously, the statistical average and expectation values are still on the list of parameters which are mostly desired during MD runs. It is well-known that the thermodynamical characteristics can be easily derived from the radial distribution function, $g(r)$, based on the fundamental relation for the system’s total energy. For example, an ideal gas would be $E_{\text{total}} = 3/2Nk_B T + 2\rho N \int r^2 V(r) g(r) dr$, where $V(r)$ is the pair potential and $g(r)$ is the radial distribution function. In more complex systems such as the water cluster, it would require more effort to split $g(r)$ into parts in terms of site–site distribution functions, e.g., $g_{\text{HH}}(r)$, $g_{\text{OO}}(r)$, and $g_{\text{HO}}(r)$, as depicted in Figure 8A.

Finally, in Figure 8 we analyze the two series of site–site distribution functions obtained from the full BOMD simulation and the M-box. This last one follows the same level of theory as the one adopted for the benzene system with DFT|PADE|TZV2P within the canonical ensemble by using a Nose–Hoover global thermostat for the 4 ps time interval. The equilibration was performed by classical AMBER fields and the TIP3P water model.^{41,42} The tasks ran for 28 water molecules in a gas cluster, as displayed in Figure 8B,C and in a nonperiodic 8000 Å³ cube. Under normal conditions with 10⁶ Pa and 300 K, it fills a space of 1000 Å³. This estimation can be derived from the equation of state for the ideal gas, $n = (pV/R)N_A$. By increasing the size of the system, we artificially introduce additional degrees of freedom which can be suppressed in the periodical solution. The M-box scheme regions are shown in Figure 8D (1, red; 2, yellow; 3, blue transparent). The complete problem depends on the force vector, which is

$\vec{F}_{\text{tot}} = \vec{F}^{1-2} + \vec{F}^{2-3} + \vec{F}^{1-3}$. As can be seen from Figure 8A, all site–site distribution functions are identical, and the integral values (e.g., $\int g_{ij}^{\text{BOMD}}(r) dr - \int g_{ij}^{\text{MBOX}}(r) dr$) are precisely zero for all ij from {O,H}. This means that the M-box model is adequate for modeling thermodynamical values based on the MD integration, which follows the M-model.

3.4. Multibox and Complex Cases. A key question that might arise is why would we need the M-box splitting scheme since the linear scaling methods have been established for decades in the DFT model? First of all, linearity does not take place in every DFT model. It is well-known that the applicability of DFT for certain problems needs to be certified, and the adopted functional approximation must be tested. For instance, pure LSD or LDA approximations are not suitable for treating d- or f-chemical elements where long-range electronic effects are essential. Even if such a system consists of a single molecule, those functionals have difficulty in correctly describing of d or f compounds. It is worth mentioning that when we say “long-range electronic interaction correlation”, it does not mean that such influence extends over the entire system (~10 nm). In some cases, it can be a very local interaction limited to 1–2 neighbors for one molecular site. The molecular system itself is a difficult problem for DFT. One well-known example is vdW complexes, which are hard to describe with chemically relevant accuracy within a pure DFT model. Additionally, natural SCF diagonalization limitation can be an issue for rather complex systems such as degenerate states, energy gap and convergence problems, etc. Second, a nonextrapolated SCF procedure could be very problematic when implemented for large systems, and at the same time good DIIS algorithms can have problems finding and verifying the real ground state.

Also, the issue of the diagonalization of large matrices might lead to problems. Rapid convergence techniques such as Lanczos and analogues⁵⁴ have strong numerical instabilities. In particular, for electronic structures, this problem grows considerably, and it is also complicated to obtain excited states. These points can partially explain certain limitations that such well established codes and models might present. It seems quite contradictory: whenever “fast models” can be employed, dynamic knowledge is hardly needed; whenever dynamic knowledge plays a role, “fast models” can hardly be used. Let us exemplify this point with an artificial molecular

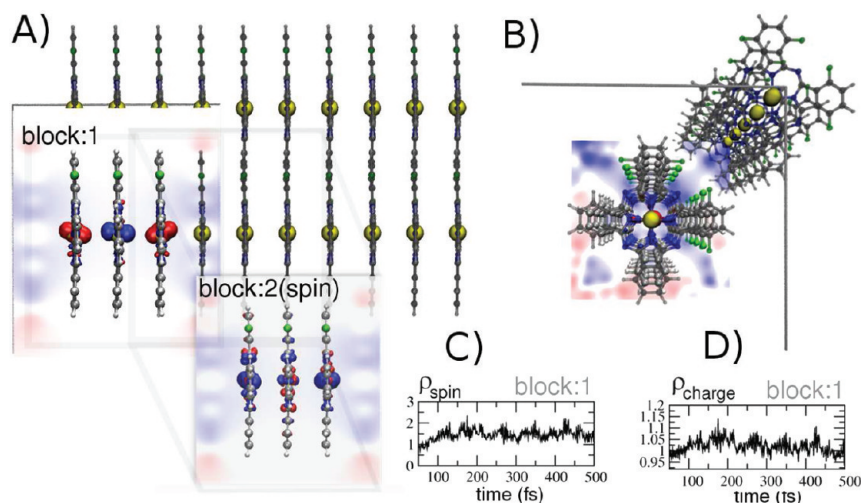


Figure 9. Equilibrium geometry of a model system of a modified Co-porphyrin stack in two projections with a profile of two blocks on top (A, B). In profile, the spin density destruction is shown for different respective blocks and the (-0.07 eV; 0.07 eV) electrostatic map given by the environment (A, B). Insets below show a short time evolution of Mulliken spin/charge density in the first block for central Co (C, D).

magnet system. We assume the system shown in Figure 9, consisting of a Co-porphyrin class 1D stack. By changing two of the hydrogen atoms in the benzene ring to fluorine atoms, an internal polarization is induced. In this case, this system might be of interest for spintronic applications since rotational operations could be used to induce such polarizations.

Even if our interest is just to obtain the equilibrium geometry of the Co-based material, some default or even more sophisticated SCF procedures cannot guarantee good results. The computational cost for 10 cobalt atoms in a line is a crucial issue for SCF convergence. We might also be interested in more interesting features such as the time spin-density dynamics recently discussed by Popov et al.⁵⁵ For this BO spin-dynamic, a natural real-space length limitation exists, and it is given by an expected tunneling length of about 3–4 localized sites. Beyond this characteristic size, single particle models hardly make sense for chemically relevant electronic structure predictions so that a local M-box treatment is a good approximation for on-site behavior. Also, forces in this system along the whole MD trajectory can be defined from the M-box model as they are also defined locally.

By using the SCF option, which is suitable for the single molecule case, we were not able to obtain a reliable wave function for whole system stack convergence (this conclusion is based on the topology of the frontal orbitals). On the basis of the Hirschfield charge prediction obtained for isolated molecules, the original electrostatic map was constructed for the QM/MM task in each block. All charges have been assumed to be equal at all MD points of the MM shells. For this task, the level of theory for the prediction of the local gradients was similar to the one used for the benzene chain problem. It means that within the each block our DFT model is also linear according to GPW ideas.^{36,37,41} For this case, the M-box splitting procedure is shown in Figure 9. As one can see, blocks 1 and 2 (three molecules each) share only one coupling molecule. The electrostatic profile volume slice in the range of (-0.07 eV, 0.07 eV) is also shown at the equilibrium point.

As an illustration for possibilities in Figure 9C and D, we present the short time fluctuations (~ 500 fs) for the spin and charge densities (Mulliken analysis for central Co in the first M-box block). This ability is obviously essential, for instance, in the electronics of complex molecular systems: small changes in the geometry

can drastically affect spin/charge distributions, and dynamic effects are also important, as proposed by Popov et al.⁵⁵

We highlight that a deeper analysis of these system dynamics goes far beyond the original scope of this manuscript. For a final remark, let us speculatively assume that the M-box local wave function which is defined in the local basis over all coupled fragments can represent in some sense an approximation for the whole system wave function in analogy to the force description. This idea once more goes back to the origin of these model FMO and FMM methods in such a way that its reconstruction does not seem to be unusual. For example, one can see a straightforward way of reconstructing such a function for the benzene chain splitting shown in Figure 2B, by using only odd blocks. This wave function, Φ_{system} built upon the combination $\{\Psi_{B_1}^{1-2}, \Psi_{B_3}^{3-4}, \dots, \Psi_{B_9}^{9-10}\}$, is an *initial guess* or an *approximation* for the whole system in the same basis set as it is for blocks. A wave function defined in this way (only odd blocks) contains vectors which are normalized and mutually orthogonal. However, since its splitting-map does have some orthogonal components, the system wave function construction requests only an additional orthogonalization step. In other words, a system wave function can be glued (as we did for forces) with respect to the each assumed splitting scheme and coupling range in a linear/nonlinear fashion depending on necessity.

By using M-box ideas, we can obtain space- and time-consistent sampling of the potential energy surface for the system with experimentally relevant sizes and chemically accurate geometrical prediction. Such approximations for the dynamical wavefunction and system Hamiltonian are suitable for physical phenomena such as those addressed here^{4–6} or even more complex structures.⁵⁶ In the future, our next step is to generalize the M-box model for the situation where system blocks can act on a potential energy surface with different multiplicity/charge sampling. This would make the local block forces couple on the electronic degrees of freedom, with the possibility of sampling along MD trajectories becoming possible.

3.5. Multibox Possible Extensions. In summary, we would like to discuss further some technical details regarding the multibox method explicitly:

- Importantly, there are no strict limitations on the level of theory for the QM (it can be, for example, some post-HF theory) region nor for MM potentials.

- (b) It is straightforward to notice that, in the M-box model, the energies and forces of each subsystem are evaluated in the presence of the surrounding MM potentials. Therefore, classical dispersion interactions for each quantum block can be included in analogy to a DFT-D model.⁴⁷ Again, DFT-D is given as an example, if a DFT model is used for the QM region.
- (c) As for all QM/MM subsystems, S_i can be computed completely independently, and it is possible to build a distribution of S_i on subgroups of processors. Because of this fact, the method allows for an almost perfect parallelization for increasing system sizes. The only peculiar aspect of this type of parallelization will be the communication of the forces, which should take a negligible amount of time and was previously implemented by CP2k developers.³⁷
- (d) If for any reason the wave function of the whole system matters, as was shown for forces, a quick and efficient guess can be constructed in a local basis set. Coefficients can be defined as a combination (depending on a splitting scheme) with respect to the blocks in which they are.
- (e) In addition to the QM/MM interactions, a block-to-block electronic structure correction might be included according to the Fedorov–Morokuma decomposition scheme. However, this correction will not be crucial for the local BO gradients on each subsystem.
- (f) The M-box includes an additional useful tool. As one can see, the M-box splitting can obtain a quick and accurate initial wave function guess, where it is possible to control its main features such as net charge or spin for a fragment of the whole system. We plan to discuss this option elsewhere. So far, it is clear from the M-box problem construction that the M-box model is partially acting on a different QM space, with real space (atomic) coupling in the local basis set, and the description can be projected to the original atomic space in the same fashion as was done for the forces in the model discussed here. A simple example: suppose that the fragment B_1 in Figure 2B has an additional replica with additional charge or another excitation. Final forces can include this information. We can continuously assign an amount of charge or excitation contributions to the fragment during the whole system dynamics again by controlling fragment forces with additional knowledge about an approximation to the multiconfiguration wave function.

4. CONCLUSIONS AND OUTLOOK

In summary, we conclude that the proposed M-box scheme is a convenient technique for molecular dynamic studies of multi-molecular systems. This model produces linear scaling with the number of submolecular blocks of the system, $\sim O(N_{\text{bl}})$. For a simple example, such as the presented one (molecular chain or covalently bonded systems), we have demonstrated a close correspondence between full *ab initio* molecular dynamics and M-box simulations. It was shown that it is possible to describe the time-dependent box crossing process and the thermodynamical characteristic for the weakly interacting systems. The correlations of the trajectories on time and space domains were also addressed, confirming that the M-box scheme is an appropriate approximation for the space–time-resolved transport phenomena studies (like energy/charge transport/transfer). We expect that employing such an accurate methodology in structural predictions can

render well-defined local electronic properties for future time-dependent transport models in large-scale systems.

AUTHOR INFORMATION

Corresponding Author

*E-mail: s.avdoshenko@gmail.com.

ACKNOWLEDGMENT

S.A. is thankful for financial support from the Erasmus Mundus programme External Co-operation (EM ECW-L04 TUD 08-11) and Dr. F. Schifmann (UCL, London), Dr. A. Popov (IFW, Dresden), Dr. I. Ioffe (MSU, Moscow), Dr. C. Toher (TU, Dresden), and Dr. C. G. Rocha (UJ, Finland) for helpful discussions.

REFERENCES

- (1) Kendrew, J. C.; Bodo, C.; Dintzis, H. M.; Parrish, R. G.; Wyckoff, H.; Phillips, D. C. The Amino-Acid Sequence X-ray Methods, and Its Correlation With Chemical Data. *Nature* **1958**, *181*, 662–666.
- (2) Lundstrom, D. Structural Genomics For Membrane Proteins. *Cell. Mol. Life Sci.* **2006**, *63*, 2597–2607.
- (3) Blundell, T. L.; Johnson, L. N. *Protein Crystallography*; Academic Press: London, 1976.
- (4) Grozema, F. C.; Berlin, Y. A.; Siebbeles, L. D. A. Mechanism of Charge Migration through DNA: Molecular Wire Behavior, Single-Step Tunneling or Hopping? *J. Am. Chem. Soc.* **2000**, *122*, 10903–10909.
- (5) Berlin, Y. A.; Burin, A. L.; Siebbeles, L. D. A.; Ratner, M. A. Conformationally Gated Rate Processes in Biological Macromolecules. *J. Phys. Chem. A* **2001**, *105*, S666–S678.
- (6) Woiczikowski, P. B.; Kubar, T.; Elstner, M.; Cuniberti, G.; Gutierrez, R.; Caetano, R. Charge Transport Through Biomolecular Wires in a Solvent: Bridging Molecular Dynamics and Model Hamiltonian Approaches. *Phys. Rev. Lett.* **2009**, *102*, 208102.
- (7) Iannuzzia, M.; Hutter, J. Comparative study of the nature of chemical bonding of corrugated graphene on Ru(0001) and Rh(111) by electronic structure calculations. *Surf. Sci.* **2011**, *15*, 1360–1368.
- (8) Ding, Y.; Iannuzzi, M.; Hutter, J. Investigation of h-BN/Rh(111) nanomesh interacting with water and atomic hydrogen. *CHIMIA* **2011**, *65*, 256–259.
- (9) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and testing of a general amber force field. *J. Comput. Chem.* **2004**, *25*, 1157–1174.
- (10) Sulpizi, M.; Raugei, S.; VandeVondele, J.; Carloni, P.; Sprik, M. Calculation of Redox Properties: AAL Understanding Short- and Long-Range Effects in Rubredoxin. *J. Phys. Chem. B* **2007**, *111*, 3969–3976.
- (11) Van Voorhis, T.; et al. The Diabatic Picture of Electron Transfer, Reaction Barriers, and Molecular Dynamics. *Annu. Rev. Phys. Chem.* **2010**, *61*, 149–170.
- (12) Gamiz-Hernandez, A. P.; Kieseritzky, G.; Ishikita, H.; Knapp, E. W. Rubredoxin Function: Redox Behavior from Electrostatics. *J. Chem. Theory Comput.* **2011**, *7*, 742–752.
- (13) Talbot, J.; Frenkel, D. Short-time correlations in liquids: Molecular-dynamics simulation of hard spheroids. *Phys. Rev. Lett.* **1990**, *65*, 2828.
- (14) Pagonbarraga, I.; Hagen, M. H. J.; Lowe, C. P.; Frenkel, D. Short-time dynamics of colloidal suspensions in confined geometries. *Phys. Rev. E* **1999**, *59*, 4458.
- (15) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. P. Development and Use of Quantum Mechanical Molecular Models. AM1: A New General Purpose Quantum Mechanical Molecular Model. *J. Am. Chem. Soc.* **1985**, *107*, 3002.
- (16) Elstner, M.; Porezag, D.; Jungnickel, G.; Elsner, J.; Haugk, M.; Frauenheim, T.; Suhai, S.; Seifert, G. Self-Consistent-Charge Density-Functional Tight-Binding Method For Simulations of Complex Materials Properties. *Phys. Rev. B* **1998**, *58*, 7260.

- (17) Greengard, L.; Rokhlin, V. A fast algorithm for particle simulations. *J. Comput. Phys.* **1987**, *73*, 325.
- (18) Clare, B. W.; Kepert, D. L. Structures, stabilities and isomerism in $C_{60}H_{36}$ and $C_{60}F_{36}$. A comparison of the AM1 Hamiltonian and density functional techniques. *THEOCHEM* **2002**, 195–207.
- (19) Cybulski, S. M.; Bledson, T. M.; Toczylowski, R. R. Comment on “Hydrogen bonding and stacking interactions of nucleic acid base pairs: A density-functional-theory treatment. *J. Chem. Phys.* **2002**, *116*, 11039.
- (20) Witek, H. A.; Irle, S.; Zheng, G.; de Jong, W. A.; Morokuma, K. Modeling carbon nanostructures with the self-consistent charge density-functional tight-binding method: Vibrational spectra and electronic structure of C_{28} , C_{60} , and C_{70} . *J. Chem. Phys.* **2006**, *125*, 214706.
- (21) Zheng, G.; et al. Parameter Calibration of Transition-Metal Elements for the Spin-Polarized Self-Consistent-Charge Density-Functional Tight-Binding (DFTB) Method: Al , Sc, Ti, Fe, Co, and Ni. *J. Chem. Theory Comput.* **2007**, *3*, 1349–1367.
- (22) Scuseria, G. E. Linear Scaling Density Functional Calculations with Gaussian Orbitals. *J. Phys. Chem. A* **1999**, *103*, 4782–4790.
- (23) Kudin, K. N.; Scuseria, G. E. Linear Scaling Density Functional Theory with Gaussian Orbitals and Periodic Boundary Conditions: Efficient Evaluation of Energy and Forces via the Fast Multipole Method. *Phys. Rev. B* **2000**, *61*, 16440–16453.
- (24) Lippert, G.; Hutter, J.; Parrinello, M. The Gaussian and Augmented-Plane-Wave Density Functional Method for *Ab Initio* Molecular Dynamics Simulations. *Theor. Chem. Acc.* **1999**, *103*, 124–140.
- (25) Guidon, M.; Schiffmann, F.; Hutter, J.; VandeVondele, F. *Ab Initio* Molecular Dynamics Using Hybrid Density Functionals. *J. Chem. Phys.* **2008**, 128.
- (26) Gao, J. Toward a Molecular Orbital Derived Empirical Potential for Liquid Simulations. *J. Phys. Rev. B.* **1997**, *101*, 657–663.
- (27) Gao, J. A molecular-orbital derived polarization potential for liquid water. *J. Phys. Rev.* **1998**, *109*, 2346.
- (28) Song, L.; Han, J.; Lin, Y.-I.; Xie, W.; Gao, J. Explicit Polarization (X-Pol) Potential Using *ab Initio* Molecular Orbital Theory and Density Functional Theory. *J. Phys. Rev. A.* **2009**, *113*, 11656–11664.
- (29) Kitaura, K.; Morokuma, K. New Energy Decomposition Scheme for Molecular-Interactions Within Hartree-Fock Approximation. *Int. J. Quantum Chem.* **1976**, *10*, 325.
- (30) Kitaura, K.; Ikeo, E.; Asada, T.; Nakano, T.; Uebayasi, M. Fragment molecular orbital method: application to polypeptides. *J. Phys. Chem. A* **2007**, *111*, 6904–6914.
- (31) Fedorov, D.; Kitaura, K. *The Fragment Molecular Orbital Method*; CRC Press: Boca Raton, FL, 2009.
- (32) Warshel, A.; Bromberg, A. Deuterium in the Initiation Step of the Thermal Reaction With Oxygen. *J. Chem. Phys.* **1970**, *52*, 1262.
- (33) Methods and Applications of Combined Quantum Mechanical and Molecular Mechanical Potentials. In *Reviews in Computational Chemistry*; Eds., Lipkowitz, K. B., Boyd, D. B., Eds.; VCH Publishers: New York, 1995; Vol. 7, pp 119–185.
- (34) Altoe, P.; Stenta, M.; Bottoni, A.; Garavelli, M. A Tunable QM/MM Approach to Chemical Reactivity, Structure and Physico-Chemical Properties Prediction. *Theor. Chem. Acc.* **2007**, *118*, 219.
- (35) Senn, H. M.; Thiel, W. QM/MM Methods for Biological Systems. *Angew. Chem., Int. Ed.* **2009**, *48*, 1198–1229.
- (36) VandeVondele, J.; Krack, M.; Mohamed, F.; Parrinello, M.; Chassaing, T.; Hutter, J. Quickstep: Fast and Accurate Density Functional Calculations Using a Mixed Gaussian and Plane Waves Approach. *Comput. Phys. Commun.* **2005**, *167*, 103–128.
- (37) The CP2K developers group (2010). <http://cp2k.berlios.de> (accessed October 2011).
- (38) Laino, T.; Mohamed, F.; Laio, A.; Parrinello, M. An efficient real space multigrid QM/MM electrostatic coupling. *J. Chem. Theory Comput.* **2005**, *6*, 1176–1184.
- (39) Laino, T.; Mohamed, F.; Laio, A.; Parrinello, M. An efficient linear-scaling electrostatic coupling for treating periodic boundary conditions in QM/MM simulations. *J. Chem. Theory Comput.* **2006**, *5*, 1370–1378.
- (40) Humphrey, W.; Dalke, A.; Schulte, K. VMD: Visual Molecular Dynamics. *J. Mol. Graphics* **1996**, *14*, 33–38.
- (41) Goedecker, S.; Teter, M.; Hutter, J. Separable Dual-Space Gaussian Pseudopotentials. *J. Phys. Rev. B.* **1996**, *54*, 1703.
- (42) Case, D. A.; Cheatham, T. A.; et al. *AMBER 10*; San Francisco, CA, 2008.
- (43) Pearlman, D. A.; Case, D. A.; Caldwell, J. W.; Ross, W. S.; Cheatham; et al. *Amber. Comp. Phys. Commun.* **1995**, *91*, 1112.
- (44) Case, D. A.; Cheatham, T.; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. *Amber. J. Comput. Chem.* **1995**, *26*, 1668.
- (45) Linstrom, P. J.; Mallard, W. G. *NIST ChemistryWebBook*, NIST Standard Reference Database 69; 2010.
- (46) Leimkuhler, B.; Reich, S. Symplectic Integration Of Constrained Hamiltonian Systems. *Math. Comput.* **1994**, *63*, 589–605 and references therein.
- (47) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A Consistent and Accurate *Ab initio* Parametrization of Density Functional Dispersion Correction (DFT-D) for the 94 Elements H-Pu. *J. Chem. Phys.* **2010**, *132*, 154104.
- (48) Huckel, E. *Z. Phys.* **1932**, *70* (3/4), 204–86. *Z. Phys.* **1932**, *72*, 310–37. *Z. Phys.* **1932**, *76*, 628–48.
- (49) Piaggio, P.; Dellepiane, G.; Piseri, L.; Tubino, R.; Taliani, C. Polarized infrared spectra of highly oriented polyacetylenes. *Sol. Commun.* **1984**, 947–956.
- (50) Head-Gordon, M. Quantum Chemistry and Molecular Processes. *J. Phys. Rev.* **1996**, *100*, 13213–13225.
- (51) Cohen, A. J.; Mori-Sanchez, P.; Yang, W. Insights into Current Limitations of Density Functional Theory. *Science* **2008**, *321*, 792–794.
- (52) Buló, R. E.; Ensing, B.; Sikkema, J.; Lucas Visscher, L. Toward a Practical Method for Adaptive QM/MM Simulations. *J. Chem. Theory Comput.* **2009**, *5*, 2212–2221.
- (53) Heyden, A.; Truhlar, D. G. Conservative Algorithm for an Adaptive Change of Resolution in Mixed Atomistic-Coarse-Grained Multiscale Simulations. *J. Chem. Theory Comput.* **2008**, 217–221.
- (54) Lanczos, C. *J. Res. Natl. Bur. Stand.* **1950**, *45*, 255. Gagliano, E.; et al. *Phys. Rev. B* **1986**, *34*, 1677.
- (55) Popov, A. A.; Chen, C.; Yang, S.; Lipps, F.; Dunsch, L. Spin-Flow Vibrational Spectroscopy of Molecules with Flexible Spin Density: Electrochemistry, ESR, Cluster and Spin Dynamics, and Bonding in $\text{TiSc}_2\text{N}@C_{80}$. *ACS Nano* **2010**, *8*, 4857–4871.
- (56) Lee, M. H.; Avdoshenko, S.; Gutierrez, R.; Cuniberti, G. Charge migration through DNA molecules in the presence of mismatches. *Phys. Rev. B* **2010**, *82*, 155455.

On the Use of Enveloping Distribution Sampling (EDS) to Compute Free Enthalpy Differences between Different Conformational States of Molecules: Application to 3_{10} -, α -, and π -Helices

Zhixiong Lin,^{†,‡} Haiyan Liu,[‡] Sereina Riniker,[†] and Wilfred F. van Gunsteren^{*,†}

[†]Laboratory of Physical Chemistry, Swiss Federal Institute of Technology, ETH, 8093 Zürich, Switzerland

[‡]School of Life Sciences and Hefei National Laboratory for Physical Sciences at the Microscale, University of Science and Technology of China (USTC), Hefei, Anhui 230027, People's Republic of China

S Supporting Information

ABSTRACT: Enveloping distribution sampling (EDS) is a powerful method to compute relative free energies from simulation. So far, the EDS method has only been applied to alchemical free energy differences, i.e., between different Hamiltonians defining different systems, and not yet to obtain free energy differences between different conformations or conformational states of a system. In this article, we extend the EDS formalism such that it can be applied to compute free energy differences of different conformations and apply it to compute the relative free enthalpy ΔG of 3_{10} -, α -, and π -helices of an alanine deca-peptide in explicit water solvent. The resulting ΔG values are compared to those obtained by standard thermodynamic integration (TI) and from so-called end-state simulations. A TI simulation requires the definition of a λ -dependent pathway which in the present case is based on hydrogen bonds of the different helical conformations. The values of $\langle(\partial V_{\text{TI}})/(\partial \lambda)\rangle_{\lambda}$ show a sharp change for a particular range of λ values, which is indicative of an energy barrier along the pathway, which lowers the accuracy of the resulting ΔG value. In contrast, in a two-state EDS simulation, an unphysical reference-state Hamiltonian which connects the parts of conformational space that are relevant to the different end states is constructed automatically; that is, no pathway needs to be defined. In the simulation using this reference state, both helices were sampled, and many transitions between them occurred, thus ensuring the accuracy of the resulting free enthalpy difference. According to the EDS simulations, the free enthalpy differences of the π -helix and the 3_{10} -helix versus the α -helix are 5 kJ mol⁻¹ and 47 kJ mol⁻¹, respectively, for an alanine deca-peptide in explicit SPC water solvent using the GROMOS 53A6 force field. The EDS method, which is a particular form of umbrella sampling, is thus applicable to compute free energy differences between conformational states as well as between systems and has definite advantages over the traditional TI and umbrella sampling methods to compute relative free energies.

1. INTRODUCTION

Calculation of free energy differences is fundamental in order to understand the properties of physical, chemical, and biological systems and phenomena. Consequently, it has for a long time been one of the central tasks of molecular simulation. Molecular dynamics (MD) simulations are widely used for such calculations.^{1–12} The ability to calculate relative free energies from MD simulations not only allows one to understand the underlying processes at the atomic level but also to probe unstable states of a system that are experimentally not accessible. Applications of free energy simulations include processes such as solvation,^{13,14} ligand binding,^{15,16} peptide/protein folding,^{17,18} ion transport,¹⁹ and so on. In order to obtain accurate free energy differences, two main challenges have to be met. First, a model for the system has to be made, i.e., a Hamiltonian or force field, that must correctly describe the thermodynamic behavior of the system. Second, an efficient scheme to calculate relative free energies from a configurational ensemble has to be found. In this article, we will concentrate on the second challenge.

Free energy differences in molecular systems can be classified into two main categories.²⁰ One is the free energy difference between different conformations or conformational states of a molecule. Typical examples include the evaluation of the free

energies of folding, of different helical and hairpin structures,^{21,22} and of transportation of ions over membranes. The other category comprises the alchemical free energy difference between different Hamiltonians, e.g., different molecules or molecular systems. Typical examples include the evaluation of relative free energies of solvation and binding.

There exists a great variety of methods for calculating relative free energies.^{1,2,4,5,8,10–12} The traditional ones for the calculation of relative conformational free energies are direct counting of the different conformations in an unbiased simulation and umbrella sampling²³ using one or more biasing potential energy functions that enhance the sampling of particular conformations. In the counting method, the free energy is calculated directly from the ratio of the numbers of sampled configurations belonging to each state. Thus, it fails if one of the conformational states is insufficiently sampled, i.e., is not stable or if there are high barriers between them. The umbrella sampling method, on the other hand, relies on the choice of a pathway and the use of a biasing potential energy term in the Hamiltonian that forces the sampling going from one state to the other. Some recently

Received: September 5, 2011

Published: October 28, 2011

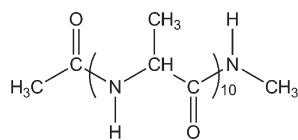


Figure 1. Chemical formula of the alanine deca-peptide studied: acetyl-(Ala)₁₀-N-methyl.

developed and sophisticated methods include local elevation umbrella sampling (LEUS)²⁴ and ball-and-stick LEUS.²⁰

In regard to alchemical perturbations, the established methods are thermodynamic integration²⁵ (TI) and free energy perturbation²⁶ (FEP). The Hamiltonians of the two different states are connected by the introduction of a coupling parameter λ . Thus, a pathway $V(\lambda)$ is defined in order to sample from one state to the other. In the TI method, the quantity $(\partial V)/(\partial \lambda)$ is averaged and integrated along the pathway. In the multistep FEP method, independent simulations at different λ values are performed, and subsequently exponential averaging is used to determine the free energy differences between neighboring λ values, which are added up to give the total free energy difference. The limitations of these two methods are, first, that a significant amount of simulation time is spent on the noninteresting (generally unphysical) intermediate states along the pathway in order to obtain enough overlap of the sampled phase space between the two states of interest. Second, the choice of a pathway $V(\lambda)$ which allows optimal sampling and convergence of the averages for each value of λ is not trivial. Therefore, some methods avoid the definition of a pathway of intermediate states and only rely on sampling of the two end states. For example, one may take the crossing point of the energy difference distributions as an estimation of the free energy difference,^{27,28} or the mean of the FEP results obtained from the two end states in the two directions, or Bennett's acceptance ratio method,^{29,30} or more generalized overlap sampling methods.^{31,32} In these methods, no pathway needs to be defined, but the overlap of the phase spaces relevant to two end states is required. Third, these methods are limited to two-state problems. For this reason, FEP has been generalized to be able to calculate relative free energies of multiple states from a single simulation of a possibly unphysical reference Hamiltonian in the so-called "one-step perturbation" method.^{16,33} Thus, no pathway has to be defined as well. However, one-step perturbation will fail if there is not enough overlap of the sampled phase space of the reference state and that of the end states. This led to the development of the method of enveloping distribution sampling (EDS).^{34–39} In the EDS method, the parameters of an unphysical reference-state Hamiltonian are iteratively optimized such that the different end states are most uniformly sampled in a single simulation, thus ensuring the accuracy of the resulting relative free energies. Using EDS, no pathway of intermediate states, e.g., a $V(\lambda)$, needs to be chosen; it is defined through the general form of the EDS Hamiltonian, which has a functional form that is solely defined by the functional forms of the end-state Hamiltonians in combination with two parameters, the smoothness parameter s_{BA} , and the reference energy offset difference ΔE_{BA}^R per pair of end states A and B , and through the procedure to optimize the s_{BA} and ΔE_{BA}^R parameters. No λ dependence of the Hamiltonian needs to be specified, only end-state Hamiltonians. In principle, EDS can be applied to multiple end-state problems, and to the case where there is no overlap between the end states.³⁵

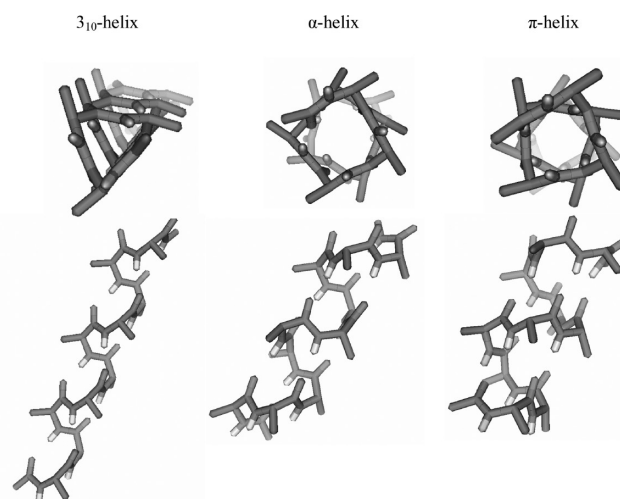


Figure 2. Top and side views of the three helices of the alanine deca-peptide.

So far, the EDS method has only been applied to alchemical free energy differences^{34–39} and not yet been used to obtain the relative free energy of different conformations or conformational states. In this article, we investigate the use of EDS to obtain the free enthalpy differences between different conformations of a solute molecule, i.e., 3_{10} -, α -, and π -helices of an alanine deca-peptide (Figures 1 and 2) in aqueous solution, and compare the results with those obtained by TI using a particular $V(\lambda)$ and by calculations from end-state simulations. We only consider two two-state problems, i.e., the free enthalpy difference between the α - and the π -helix and between the α - and the 3_{10} -helix.

2. THEORY

Assume we wish to calculate the free enthalpy difference between two conformations, α and β , of a molecule, and one or both of them is not the most stable one of the molecule. We may use the EDS technique to obtain the free enthalpy difference by defining the EDS reference Hamiltonian as follows.

Two restraining energy function terms are defined which restrain the molecular conformations to conformation α or to conformation β , i.e., $V_X^{\text{rest}}(\vec{r}^N; K_X^{\text{rest}}, \vec{r}_{0\xi}^N)$ where $X = A$ or B and $\vec{r}_{0\xi}^N$ is the set of parameters which characterizes the conformation ξ , $\xi = \alpha$ or β , e.g., through particular hydrogen-bond distance ranges or torsional-angle ranges, and K_X^{rest} is the restraining force constant. Thus, the resulting Hamiltonian for the end state X is

$$V_X(\vec{r}^N) = V_X^{\text{rest}}(\vec{r}^N; k_X^{\text{rest}}, \vec{r}_{0\xi}^N) + V^{\text{phys}}(\vec{r}^N) \quad (1)$$

where $V^{\text{phys}}(\vec{r}^N)$ is the interaction function of a particular force field. Then, we may construct an EDS reference-state Hamiltonian:

$$\begin{aligned} V_R(\vec{r}^N; s, E_{BA}^R) &= -k_B T s^{-1} \ln \left\{ e^{-s(V_A^{\text{rest}}(\vec{r}^N) - E_A^R)/k_B T} \right. \\ &\quad \left. + e^{-s(V_B^{\text{rest}}(\vec{r}^N) - E_B^R)/k_B T} \right\} + V^{\text{phys}}(\vec{r}^N) \\ &= V^{\text{EDS, rest}}(\vec{r}^N; s, E_{BA}^R) + V^{\text{phys}}(\vec{r}^N) \end{aligned} \quad (2)$$

where s is a smoothness parameter and $E_B^R - E_A^R = E_{BA}^R$ is an energy offset parameter difference, which are chosen such as to optimize the sampling of both end states A and B .

In the original EDS implementation, the configurations \vec{r}^N that are sampled by the reference Hamiltonian H_R , i.e., V_R are not assigned to any conformational states. They are considered³⁴ to belong to state A if

$$(V_A^{\text{rest}}(\vec{r}^N) - E_A^R) < (V_B^{\text{rest}}(\vec{r}^N) - E_B^R) \quad (3)$$

In the case considered here, configurations must be separated into different sets, i.e., different conformational states: they belong to set α if

$$V_A^{\text{rest}}(\vec{r}^N) \leq E_\alpha^{\text{thres}} \text{ and } V_B^{\text{rest}}(\vec{r}^N) > E_\beta^{\text{thres}} \quad (4)$$

they belong to set β if

$$V_A^{\text{rest}}(\vec{r}^N) > E_\alpha^{\text{thres}} \text{ and } V_B^{\text{rest}}(\vec{r}^N) \leq E_\beta^{\text{thres}} \quad (5)$$

or they may belong to neither of them, called sets γ and δ with set γ defined by

$$V_A^{\text{rest}}(\vec{r}^N) > E_\alpha^{\text{thres}} \text{ and } V_B^{\text{rest}}(\vec{r}^N) > E_\beta^{\text{thres}} \quad (6)$$

and set δ defined by

$$V_A^{\text{rest}}(\vec{r}^N) \leq E_\alpha^{\text{thres}} \text{ and } V_B^{\text{rest}}(\vec{r}^N) \leq E_\beta^{\text{thres}} \quad (7)$$

Generally, set δ should contain no or only a few configurations in order to make a meaningful distinction between sets α and β .

Here, the configurations that belong to sets α and β are defined via an energy threshold criterion E_ξ^{thres} , which maps configurations \vec{r}^N onto an energy $V_X^{\text{rest}}(\vec{r}^N)$ using the same function $V_X^{\text{rest}}(\vec{r}^N)$ that is used in the reference Hamiltonian. This means that the configurations that belong to sets α and β are defined through eqs 4 and 5, respectively. We note that these sets α and β differ from the conformational ensembles A and B that are through the end-state Hamiltonians defined by eq 1.

Alternatively, the conformational sets α and β could be defined using a geometric measure such as an atom-positional root-mean-square deviation (RMSD) from a given configuration, either in Cartesian or in internal torsional coordinates, instead of using the restraining functions V_X^{rest} and threshold energies E_ξ^{thres} . Configurations then belong to set α if

$$\text{RMSD}(\vec{r}^N, \vec{r}_\alpha^N) \leq \text{RMSD}_\alpha^{\text{thres}} \text{ and } \text{RMSD}(\vec{r}^N, \vec{r}_\beta^N) > \text{RMSD}_\beta^{\text{thres}} \quad (8)$$

they belong to set β if

$$\text{RMSD}(\vec{r}^N, \vec{r}_\alpha^N) > \text{RMSD}_\alpha^{\text{thres}} \text{ and } \text{RMSD}(\vec{r}^N, \vec{r}_\beta^N) \leq \text{RMSD}_\beta^{\text{thres}} \quad (9)$$

or they belong to neither of them, called set γ and set δ :

$$\gamma: \text{RMSD}(\vec{r}^N, \vec{r}_\alpha^N) > \text{RMSD}_\alpha^{\text{thres}} \text{ and } \text{RMSD}(\vec{r}^N, \vec{r}_\beta^N) > \text{RMSD}_\beta^{\text{thres}} \quad (10)$$

$$\delta: \text{RMSD}(\vec{r}^N, \vec{r}_\alpha^N) \leq \text{RMSD}_\alpha^{\text{thres}} \text{ and } \text{RMSD}(\vec{r}^N, \vec{r}_\beta^N) \leq \text{RMSD}_\beta^{\text{thres}} \quad (11)$$

Again, the thresholds $\text{RMSD}_\xi^{\text{thres}}$ should be chosen such that set δ contains no or only a few configurations.

In the procedure and expressions used in the optimization of the parameters s and $E_B^R = E_{BA}^R$ (E_A^R is standardly set to zero in two-state EDS), configurations that belong to sets γ and δ can be ignored. Thus, we get for updating the energy offset

E_B^R (corresponds to eq 13 of ref 36):

$$E_B^R(\text{new}) = -k_B T \ln \left\{ e^{-(V_A^{\text{rest}} - V_B^{\text{rest}} + E_B^R(\text{old}))/k_B T} + 1 \right\}^{-1}_{R, \text{not}\gamma, \text{not}\delta} + E_B^R(\text{old}) \quad (12)$$

where configurations of sets γ and δ are excluded when calculating the ensemble average over the ensemble of the reference state R .

For updating or rather choosing a new s parameter, we calculate

$$s_A = - \left\{ \ln \left\langle e^{-(|V_B^{\text{rest}} - V_A^{\text{rest}}| - E_{BA}^R)/k_B T} \right\rangle_A \right\}^{-1} \quad (13)$$

and

$$s_B = - \left\{ \ln \left\langle e^{-(|V_A^{\text{rest}} - V_B^{\text{rest}}| + E_{BA}^R)/k_B T} \right\rangle_B \right\}^{-1} \quad (14)$$

and take the lowest s value as the new s

$$s = \min(s_A, s_B) \quad (15)$$

which corresponds to eq 14 of ref 36.

Ensembles A and B are obtained by reweighting the configurations generated using the reference state R to the corresponding end state A or B . For a quantity $Q(\vec{r}^N)$, which is a function of the coordinates \vec{r}^N , we have

$$\langle Q \rangle_X = \frac{\int Q(\vec{r}^N) e^{-V_X(\vec{r}^N)/k_B T} d\vec{r}^N}{\int e^{-V_X(\vec{r}^N)/k_B T} d\vec{r}^N} \quad (16)$$

or using the ensemble R

$$\langle Q \rangle_X = \frac{\langle Q e^{-(V_X - V_R)/k_B T} \rangle_R}{\langle e^{-(V_X - V_R)/k_B T} \rangle_R} \quad (17)$$

Subsequently, the ensemble averaging in eq 17 could be restricted to the sets α and β . In that case, these restricted ensemble averages can be written as

$$\begin{aligned} \langle Q \rangle_A &= \frac{\langle Q e^{-(V_A - V_R)/k_B T} \rangle_{R, \text{not}\gamma, \text{not}\delta}}{\langle e^{-(V_A - V_R)/k_B T} \rangle_{R, \text{not}\gamma, \text{not}\delta}} \\ &= \frac{\langle Q e^{-[V_A^{\text{rest}} - V^{\text{EDS, rest}}(s, E_{BA}^R)]/k_B T} \rangle_{R, \text{not}\gamma, \text{not}\delta}}{\langle e^{-[V_A^{\text{rest}} - V^{\text{EDS, rest}}(s, E_{BA}^R)]/k_B T} \rangle_{R, \text{not}\gamma, \text{not}\delta}} \end{aligned} \quad (18)$$

and

$$\begin{aligned} \langle Q \rangle_B &= \frac{\langle Q e^{-(V_B - V_R)/k_B T} \rangle_{R, \text{not}\gamma, \text{not}\delta}}{\langle e^{-(V_B - V_R)/k_B T} \rangle_{R, \text{not}\gamma, \text{not}\delta}} \\ &= \frac{\langle Q e^{-[V_B^{\text{rest}} - V^{\text{EDS, rest}}(s, E_{BA}^R)]/k_B T} \rangle_{R, \text{not}\gamma, \text{not}\delta}}{\langle e^{-[V_B^{\text{rest}} - V^{\text{EDS, rest}}(s, E_{BA}^R)]/k_B T} \rangle_{R, \text{not}\gamma, \text{not}\delta}} \end{aligned} \quad (19)$$

In this way, erratic irrelevant energy values due to irrelevant configurations not belonging to sets α and β are excluded from influencing the parameter optimization for sampling of sets α and β . Furthermore, configurations which belong to set δ that have low V_X^{rest} values are excluded too.

Table 1. Overview of the Simulations

end-state simulations					
simulations	initial structure	simulation time [ns]	no. of solvent molecules	$\langle V_X^{\text{rest}} \rangle$ [kJ mol ⁻¹]	av. hydrogen bonds [%]
3 ₁₀ -helix	3 ₁₀ -helix	11	4336	31.7	65
α -helix	α -helix	11	3204	3.7	59
π -helix	π -helix	11	2660	1.9	71
thermodynamic integration (TI) simulations					
simulations	initial structure	simulation time [ns]	no. of solvent molecules		
α -helix \rightarrow π -helix	α -helix	11 \times 2	3204		
α -helix/ π -helix EDS parameter update simulations					
simulations	initial structure	simulation time [ns]	no. of solvent molecules	update scheme	
update1	α -helix	128 \times 0.15	3204	update after 1, 3, 7, 13... of 0.15 ns simulations	
update2	π -helix	128 \times 0.15	2660	update after 1, 3, 7, 13... of 0.15 ns simulations	
update3	α -helix	128 \times 0.15	3204	update after 1, 3, 7, 15... of 0.15 ns simulations	
update4	π -helix	128 \times 0.15	2660	update after 1, 3, 7, 15... of 0.15 ns simulations	
update5	α -helix	128 \times 0.15	3204	update after 1, 3, 7, 15... of 0.15 ns simulations, with excluding noninteresting states	
update6	π -helix	128 \times 0.15	2660	update after 1, 3, 7, 15... of 0.15 ns simulations, with excluding noninteresting states	
α -helix/ π -helix EDS evaluation simulations					
simulations	initial structure	simulation time [ns]	no. of solvent molecules	parameters taken from	
EDS1	π -helix	51	2660	update1: $s = 0.16$, $E_B^R = 13.9$	
EDS2	π -helix	51	2660	update2: $s = 0.21$, $E_B^R = 10.1$	
EDS3	π -helix	51	2660	update3: $s = 0.30$, $E_B^R = 7.2$	
EDS5	π -helix	51	2660	update5: $s = 0.16$, $E_B^R = 14.1$	
α -helix/3 ₁₀ -helix EDS parameter update simulations					
simulations	initial structure	simulation time [ns]	no. of solvent molecules	update scheme	
update7	α -helix	128 \times 0.15	3204	update after 1, 3, 7, 15... of 0.15 ns simulations	
update8	3 ₁₀ -helix	128 \times 0.15	4336	update after 1, 3, 7, 15... of 0.15 ns simulations	
α -helix/3 ₁₀ -helix EDS evaluation simulations					
simulations	initial structure	simulation time [ns]	no. of solvent molecules	parameters taken from	
EDS7	α -helix	51	3204	update7: $s = 0.03$, $E_B^R = 68.7$	

The free enthalpy difference between two end-state Hamiltonians B and A in the EDS simulation is evaluated through³⁶

$$\Delta G_{BA} = G_B - G_A = \Delta G_{BR} - \Delta G_{AR} = -k_B T \ln \frac{\langle e^{-(H_B - H_R)/k_B T} \rangle_R}{\langle e^{-(H_A - H_R)/k_B T} \rangle_R} \quad (20)$$

The expression used to obtain the free enthalpy difference between conformational sets β and α from an ensemble generated using the reference-state Hamiltonian $V_R(\vec{r}^N; s, E_{BA}^R)$ reads

$$\Delta G_{\beta\alpha} = G_\beta - G_\alpha = -k_B T \ln \left\{ \frac{N_\beta(V^{\text{phys}})}{N_\alpha(V^{\text{phys}})} \right\} \quad (21)$$

where $N_\xi(V^{\text{phys}})$ is the number of configurations belonging to set ξ in an ensemble generated using V^{phys} . In terms of the ensemble R generated using the reference-state potential energy V_R we get

(see Appendix)

$$\Delta G_{\beta\alpha} = -k_B T \ln \left\{ \frac{\langle e^{+V^{\text{EDS,rest}}/k_B T} \rangle_{R, \text{set}\beta}}{\langle e^{+V^{\text{EDS,rest}}/k_B T} \rangle_{R, \text{set}\alpha}} \times \frac{N_\beta(V_R)}{N_\alpha(V_R)} \right\} \quad (22)$$

In other words, the ensemble R that was generated using the biasing potential energy function $V^{\text{EDS,rest}}$ is reweighted using eq 22, and the configurations of the sets α and β are used in the averaging via their relative populations in the ensemble R , i.e., $N_\alpha(V_R)$ and $N_\beta(V_R)$

$$\frac{N_\beta(V_R)}{N_\alpha(V_R)} = \frac{\langle \delta(\vec{r}^N - \vec{r}_\beta^N) \rangle_R}{\langle \delta(\vec{r}^N - \vec{r}_\alpha^N) \rangle_R} \quad (23)$$

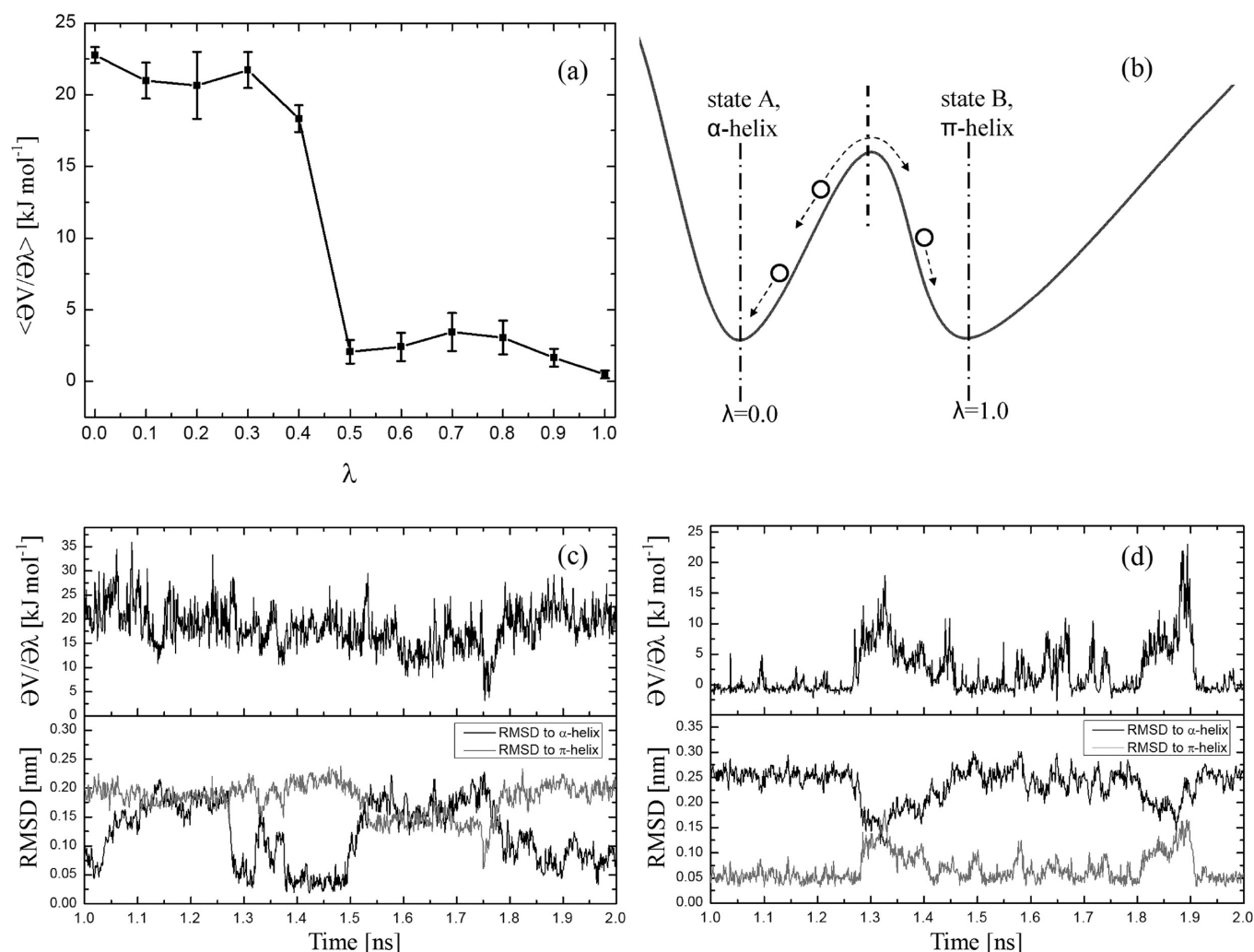


Figure 3. Thermodynamic integration (TI) simulations from state A (α -helix, $\lambda = 0.0$) to state B (π -helix, $\lambda = 1.0$) which restrain alanine deca-peptide into conformational sets α (α -helix) and β (π -helix), respectively. (a) $\langle (\partial V_{TI}) / (\partial \lambda) \rangle_\lambda$ as a function of λ . (b) A schematic representation of V^{phys} , the nonbiased potential energy at different configurations corresponding to different λ values, and the λ -dependent pathway and real energy barrier between the two states. (c, d) Time evolution of $(\partial V_{TI}) / (\partial \lambda)$ and backbone atom-positional RMSD of the peptide with respect to the α -helix or the π -helix of the TI simulations at $\lambda = 0.4$ (c) or $\lambda = 0.5$ (d).

Equation 22 can be simply rewritten as

$$\Delta G_{\beta\alpha} = -k_B T \ln \frac{N_\beta(V_R)}{N_\alpha(V_R)} - k_B T \ln \left\langle e^{+V^{EDS,rest}/k_B T} \right\rangle_{R, set\beta} + k_B T \ln \left\langle e^{+V^{EDS,rest}/k_B T} \right\rangle_{R, set\alpha} \quad (24)$$

Equation 24 is equivalent to the expression used in conformational state-specific one-step perturbation.⁴⁰ In other words, the EDS reference-state Hamiltonian can be used as the reference state in one-step perturbation, ensuring sufficient sampling of the conformational end states, which is reached by optimizing the parameters s and E_{BA}^R .

If simulations based on the end state potential energy functions $V_X(\vec{r}^N)$, see eq 1, are available, these ensembles $X = A$ and $X = B$ can also be used to obtain the free enthalpy difference

between conformational sets β and α (see Appendix):

$$\Delta G_{\beta\alpha} = -k_B T \ln \left\{ \frac{\langle e^{+V_B^{rest}/k_B T} \rangle_{B, set\beta}}{\langle e^{+V_A^{rest}/k_B T} \rangle_{A, set\alpha}} \times \frac{\langle 1 \rangle_{B, set\beta}}{\langle 1 \rangle_{A, set\alpha}} \times \frac{\langle e^{-(V_M - V_A)/k_B T} \rangle_A}{\langle e^{-(V_M - V_B)/k_B T} \rangle_B} \right\} \quad (25)$$

in which V_M is an intermediate state connecting two end states. If we use the EDS reference-state Hamiltonian as the intermediate state, the ensemble averages in the last factor of eq 25 can be written as

$$\left\langle e^{-(V_M - V_X)/k_B T} \right\rangle_X = \left\langle e^{-(V^{EDS,rest(s, E_{BA}^R)} - V_X^{rest})/k_B T} \right\rangle_X \quad (26)$$

We refer to the Appendix for the derivations of eqs 22 and 25.

3. MOLECULAR MODEL AND COMPUTATIONAL METHOD

3.1. Molecular Model, Definition of End-State Hamiltonians, and Conformational Sets. The model system considered is an

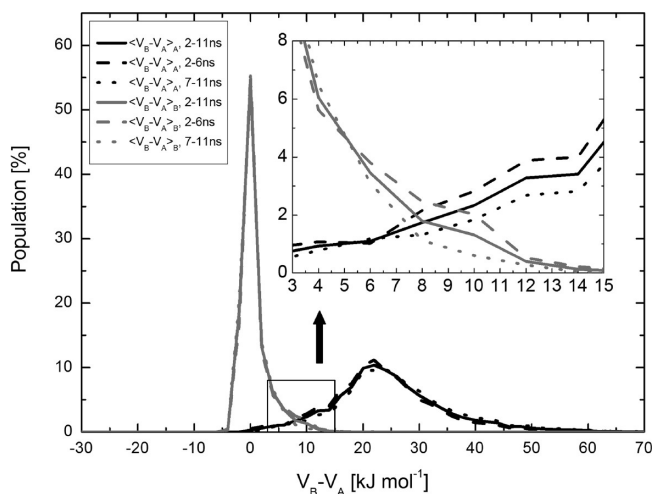


Figure 4. Distributions of the energy difference $V_B - V_A$ in the simulations of end states A (α -helix) and B (π -helix) for different simulation periods.

alanine deca-peptide capped at both termini with methyl groups, acetyl-(Ala)₁₀-N-methyl (Figure 1), solvated in water. The GRO-MOS force field 53A6⁴¹ was used for the peptide, and the water solvent molecules were represented by a rigid three-site simple-point-charge (SPC)⁴² model.

The restraining potential energy term $V_X^{\text{rest}}(\vec{r}^N; K_X^{\text{rest}}, r_{0\xi}^N)$ used to characterize different end-state Hamiltonians is defined as an attractive half-harmonic function applied to all of the hydrogen-bonding pairs of O and H atoms:

$$V_X^{\text{rest}}(\vec{r}^N; K_X^{\text{rest}}, r_{0\xi}^N) = \frac{1}{2} K_X^{\text{rest}} \sum_{i=1}^{N_{HB,\xi}} (d_{iX} - d_{0\xi})^2 \text{ when } d_{iX} > d_{0\xi} \\ = 0 \text{ when } d_{iX} \leq d_{0\xi} \quad (27)$$

where $X = A, B$, and C , which restrains the peptide into an α -, π -, and 3_{10} -helix, respectively (Figure 2). $N_{HB,\xi}$ is the number of hydrogen bonds (8, 7, or 9) of these three helices, and d_{iX} is the distance between the hydrogen bonding O and H atoms (Tables S1–S3, Supporting Information). K_X^{rest} is the force constant, and $d_{0\xi}$ is the reference distance. They were set to $30 \text{ kJ mol}^{-1} \text{ nm}^{-2}$ and 0.25 nm for state A, $150 \text{ kJ mol}^{-1} \text{ nm}^{-2}$ and 0.25 nm for state B, and $2700 \text{ kJ mol}^{-1} \text{ nm}^{-2}$ and 0.19 nm for state C. The parameters were chosen such that in the end-state simulations, the averages of corresponding helical hydrogen-bond populations are about 60–70% (Table 1 and Supporting Information Tables S1–S3 and Figures S1–S3).

The conformational sets α and β corresponding to the helices were defined through atom-positional root-mean-square deviation (RMSD) of the backbone atoms (N, CA, C) of the peptide (including the two termini) from the ideal helix, according to eqs 8 and 9. The three ideal helices (\vec{r}_{ξ}^N) were defined through the φ and ψ backbone torsional-angle values (-57.8° , -47.0°) for the α -helix, (-57.0° , -70.0°) for the π -helix, and (-49.0° , -27.0°) for the 3_{10} -helix. The RMSD threshold value $\text{RMSD}_{\xi}^{\text{thres}}$ was set to 0.15 nm for all three helices.

In the TI and EDS simulations, the peptide was solvated in different numbers of water molecules (Table 1). In the end-state simulations, different numbers of solvent molecules were used too, depending on the size of the solute; a minimum distance of 1.4 nm of any solute atom to the walls of the periodic box was required. A test simulation of the π -helix solvated in 3204

Table 2. Free Enthalpy Differences (in kJ mol^{-1}) of the Two End-State Hamiltonians B (π -helix) and A (α -helix), ΔG_{BA} , and the Two Conformational Sets β (π -helix) and α (α -helix), $\Delta G_{\pi\alpha}$ of the Alanine Deca-Peptide in Aqueous Solution^a

	ΔG_{BA}	$\Delta G_{\pi\alpha}$
TI		
TI	10.6 ± 1.1	
end-state simulations ^b		
crossing point	8.1 ± 2	
M1		3.0 ± 0.9
M2		4.1 ± 1.0
M3		4.7 ± 1.2
M4		5.3 ± 1.3
EDS ^c		
EDS1	13.7 ± 0.9	5.6 ± 0.8
EDS2	10.8 ± 0.6	4.5 ± 0.7
EDS3	13.0 ± 0.8	4.5 ± 0.8
EDS5	10.5 ± 0.6	5.6 ± 0.7

^a Statistical uncertainties (except the one for the crossing point) were estimated by block averaging.⁴⁹ ^b End-state simulations: ΔG_{BA} was calculated through the crossing point of the energy difference distributions, see also Figure 2. $\Delta G_{\pi\alpha}$ was calculated through eq 25 using EDS reference-state Hamiltonians as intermediate states. M1: $s = 0.30$, $E_B^R = 7.2 \text{ kJ mol}^{-1}$. M2: $s = 0.18$, $E_B^R = 12.0 \text{ kJ mol}^{-1}$. M3: $s = -0.30$, $E_B^R = 0.0 \text{ kJ mol}^{-1}$. M4: $s = -1.00$, $E_B^R = 0.0 \text{ kJ mol}^{-1}$. ^c EDS: ΔG_{BA} was calculated through eq 20. $\Delta G_{\pi\alpha}$ was calculated through eq 22. The EDS parameters are given in Table 1.

water molecules did not yield significantly different results (data not shown).

3.2. Thermodynamic Integration (TI). The TI simulations were carried out from state A to state B, which restrain the peptide into the conformational sets α (α -helix) and β (π -helix), respectively. The λ -dependent potential energy term $V(\lambda)$ of the Hamiltonian for the TI simulations was defined as the sum of two attractive half-harmonic functions for the end-state Hamiltonians A and B (eq 27), multiplied by $(1 - \lambda)$ and λ , respectively, added to $V^{\text{phys}}(\vec{r}^N)$:

$$V_{\text{TI}}(\vec{r}^N; \lambda) = (1 - \lambda) \frac{1}{2} K_A^{\text{rest}} \sum_{i=1}^{N_{HB,\alpha}} (d_{i\alpha} - d_{0\alpha})^2 \\ + \lambda \frac{1}{2} K_B^{\text{rest}} \sum_{i=1}^{N_{HB,\beta}} (d_{i\beta} - d_{0\beta})^2 + V^{\text{phys}}(\vec{r}^N) \quad (28)$$

where $d_{i\alpha}$ and $d_{i\beta}$ are the distances between the α -helical or π -helical hydrogen bonding atoms O and H, respectively, and the harmonic restraints are only applied when $d_{i\xi} > d_{0\xi}$. The free enthalpy difference between the two end-state Hamiltonians B and A can be calculated through

$$\Delta G_{BA} = \int_0^1 \left\langle \frac{\partial V_{\text{TI}}}{\partial \lambda} \right\rangle_{\lambda} d\lambda \quad (29)$$

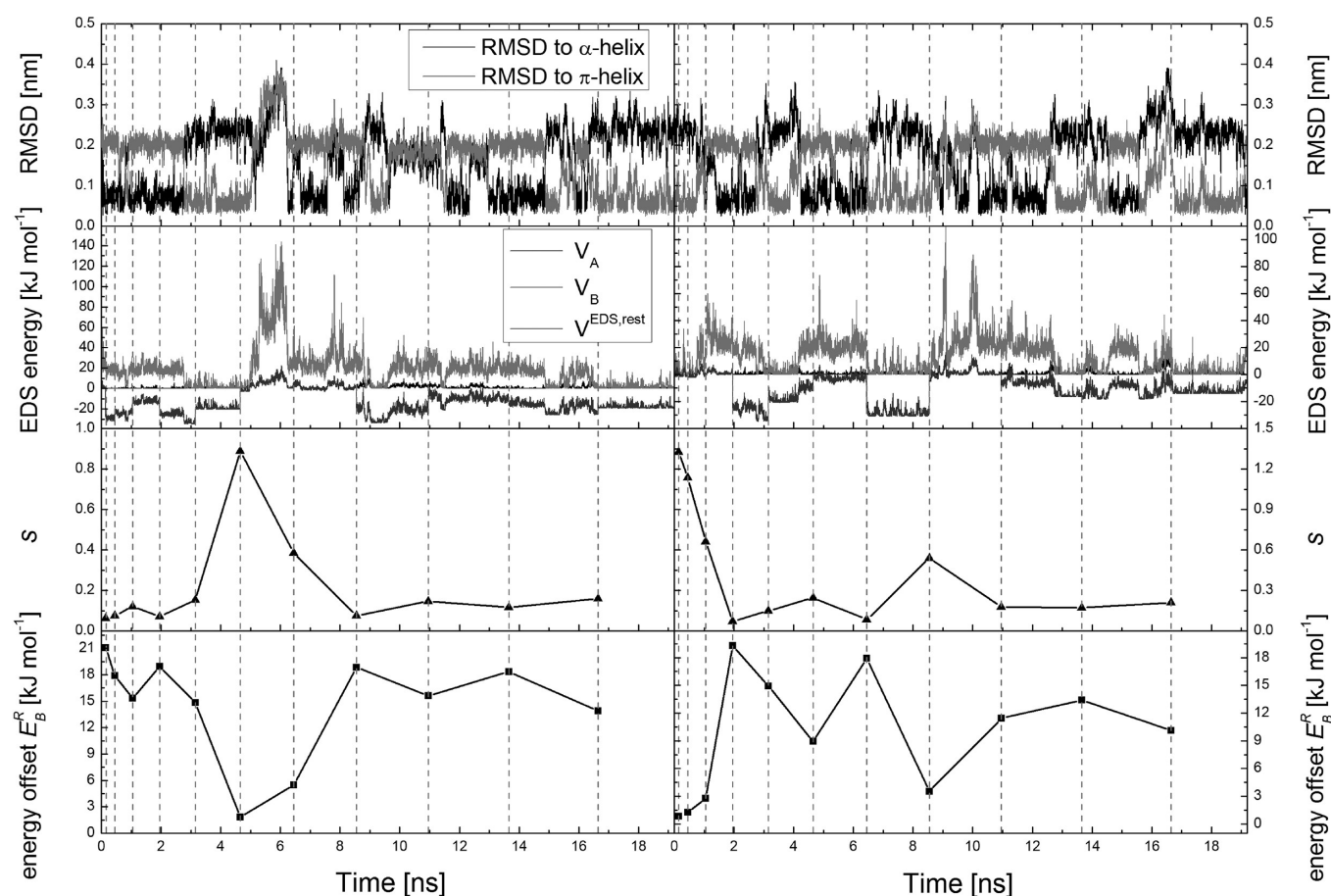


Figure 5. Time evolutions of different properties in the α -helix/ π -helix EDS parameter update simulations *update1* (left panels) and *update2* (right panels), see also Table 1. From top to bottom: backbone atom-positional RMSD of the peptide with respect to the α -helix or π -helix, the restraining and the EDS reference potential energies, smoothness parameter s , energy offset E_B^R . The vertical dashed lines show when the updates of s and E_B^R are carried out.

with

$$\frac{\partial V_{\Pi}}{\partial \lambda} = -\frac{1}{2}K_A^{\text{rest}} \sum_{i=1}^{N_{\text{HB},\alpha}} (d_{i\alpha} - d_{0\alpha})^2 + \frac{1}{2}K_B^{\text{rest}} \sum_{i=1}^{N_{\text{HB},\beta}} (d_{i\beta} - d_{0\beta})^2$$

$$\begin{aligned} \text{when } d_{i\xi} > d_{0\xi} \\ = 0 \text{ when } d_{i\xi} \leq d_{0\xi} \end{aligned} \quad (30)$$

At each of the 11 equidistant λ values, the system was equilibrated for 1 ns followed by 1 ns of production. The final configuration after 1 ns simulation at a λ value was used as the starting configuration for the next λ value.

3.3. End-State Simulations. The three end-state simulations for the three helices were carried out for 11 ns. The energy difference $V_B(\vec{r}^N) - V_A(\vec{r}^N)$ defined through eqs 1 and 27 was calculated in the simulations of the end states A and B. The crossing point of the two distributions of $V_B(\vec{r}^N) - V_A(\vec{r}^N)$ generated in the two simulations was used as an estimate for the free enthalpy difference between the end states B and A. The first 1 ns of both simulations was treated as equilibration time, and the last 10 ns were used to calculate the free enthalpy difference. In addition, the last 10 ns were divided into two continuous 5 ns periods, for which the corresponding distributions and crossing points were also calculated.

The free enthalpy difference between the conformational sets β (π -helix) and α (α -helix) was calculated through eq 25 using four different EDS reference-state Hamiltonians as the intermediate states. M1: $s = 0.30$, $E_{BA}^R = 7.2 \text{ kJ mol}^{-1}$. M2: $s = 0.18$, $E_{BA}^R = 12.0 \text{ kJ mol}^{-1}$. M3: $s = -0.30$, $E_{BA}^R = 0.0 \text{ kJ mol}^{-1}$. M4: $s = -1.00$, $E_{BA}^R = 0.0 \text{ kJ mol}^{-1}$.

3.4. Enveloping Distribution Sampling (EDS) Simulations. Six EDS parameter update simulations were performed to find the optimal reference state parameters for calculating the free enthalpy difference between the π -helix and the α -helix for $128 \times 0.15 \text{ ns}$ (Table 1). Either the α -helix or the π -helix served as the initial configuration. The parameters s and $E_B^R = E_{BA}^R$ (E_A^R is standardly set to zero in two-state EDS) were updated at fixed time points: after the first, third, seventh, and 13th etc. 0.15 ns simulation periods, or after the first, third, seventh, and 15th etc. 0.15 ns simulation periods. That is, the simulation time period was either increased by 0.3 ns or doubled after each update. The new s and E_B^R parameters were calculated without excluding any configurations or with excluding the noninteresting conformational sets γ and δ through eqs 12, 18, and 19. We refer to ref 36 for details of the updating scheme.

Four of the six resulting s and E_B^R parameters were used for EDS evaluation simulations of 51 ns (Table 1). The free enthalpy difference between the two end states, i.e. ΔG_{BA} , was calculated through eq 20, and the free enthalpy difference between the two conformational sets, i.e. $\Delta G_{\pi\omega}$ was calculated through eq 22.

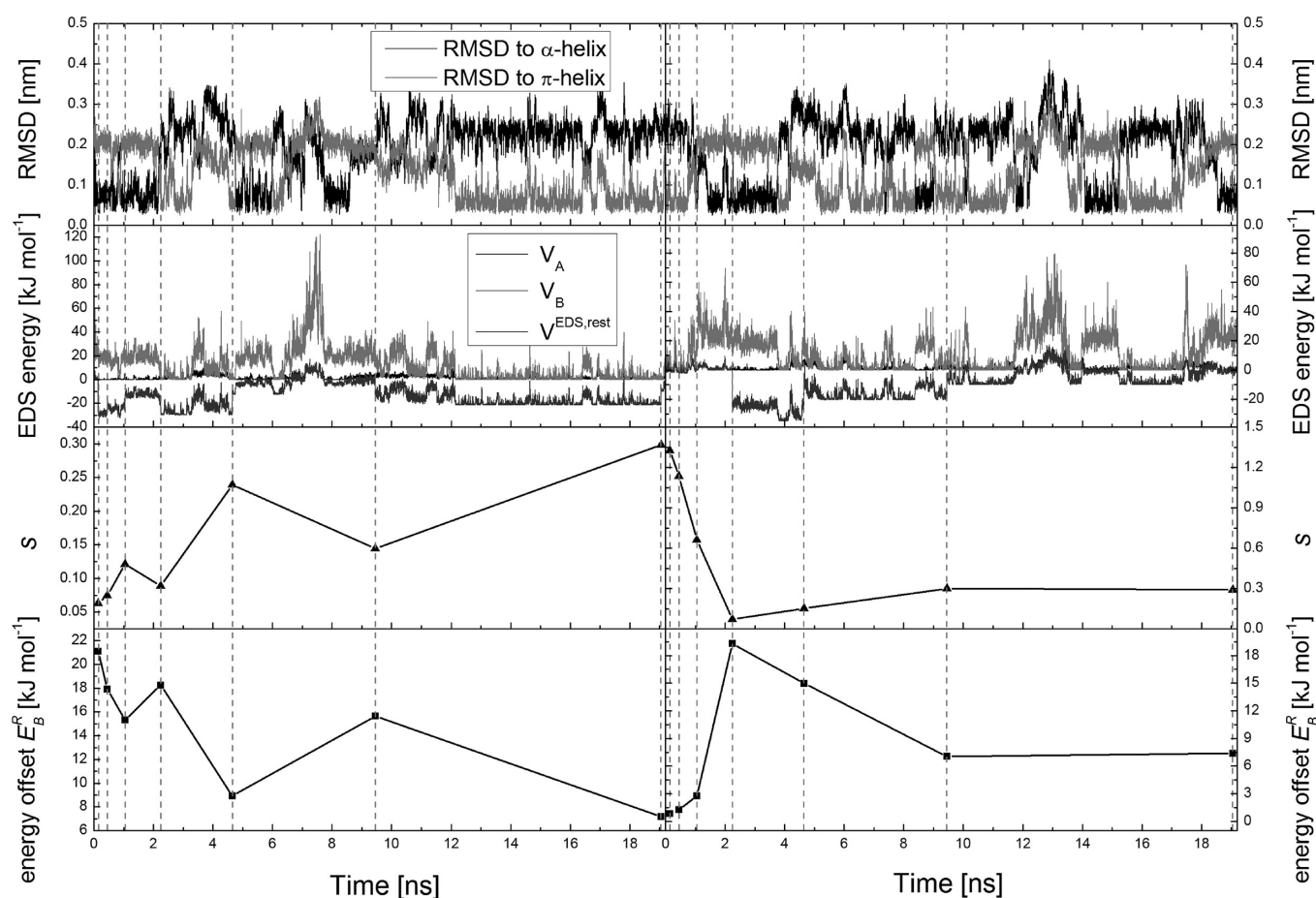


Figure 6. Time evolutions of different properties in the α -helix/ π -helix EDS parameter update simulations *update3* (left panels) and *update4* (right panels), see also Table 1. From top to bottom: backbone atom-positional RMSD of the peptide with respect to the α -helix or π -helix, the restraining and the EDS reference potential energies, smoothness parameter s , energy offset E_B^R . The vertical dashed lines show when the updates of s and E_B^R are carried out.

Two EDS parameter update simulations and one EDS evaluation simulation was carried out for calculating the free enthalpy difference between the 3_{10} -helix and the α -helix (Table 1).

3.5. Simulation Setup and Analysis. The starting configurations of the α -, π -, and 3_{10} -helices were constructed by setting all of the φ and ψ backbone torsional angles to $(-57.8^\circ, -47.0^\circ)$, $(-57.0^\circ, -70.0^\circ)$, and $(-49.0^\circ, -27.0^\circ)$, respectively. The initial configurations were first energy-minimized in a vacuum with all hydrogen bonds restrained. Each of them was then solvated in a rectangular box containing explicit SPC⁴² water. The resulting numbers of water molecules are listed in Table 1. The solvated configurations were further energy minimized to remove possible steric clashes. These configurations were used as the reference configurations for the atom-positional RMSD calculations. An equilibration scheme was carried out for each system. The atom velocities were generated from a Maxwell distribution at 60 K, and the simulation temperature was gradually raised to 300 K, while the strength of the position-restraining potential energy term for the solute atoms was decreased from $2.5 \times 10^4 \text{ kJ mol}^{-1} \text{ nm}^{-2}$ to $25 \text{ kJ mol}^{-1} \text{ nm}^{-2}$.

The simulations were carried out at a constant temperature of 300 K and a constant pressure of 1 atm using the GROMOS simulation package.^{43–45} The solute molecules and the water solvent were separately coupled to a temperature bath at 300 K

by means of weak coupling,⁴⁶ using a relaxation time of 0.1 ps. The pressure was calculated with a molecular virial and held constant by weak coupling⁴⁶ to an external pressure bath with a relaxation time of 0.5 ps, using an isothermal compressibility of $4.575 \times 10^{-4} (\text{kJ mol}^{-1} \text{ nm}^{-3})^{-1}$. All bond lengths and the geometry of the water molecules were constrained using the SHAKE algorithm⁴⁷ with a relative geometric accuracy of 10^{-4} , allowing a time step of 2 fs in the leapfrog algorithm to integrate the equations of motion. For the treatment of the nonbonded interactions, twin-range cutoff radii of 0.8/1.4 nm were used. Interactions within 0.8 nm were evaluated every time step. The intermediate range interactions were updated every fifth time step, and the long-range electrostatic interactions beyond 1.4 nm were approximated by a reaction field force⁴⁸ according to a dielectric continuum with a dielectric permittivity of 61.

Trajectory coordinates and energies were stored at 1 ps intervals for analysis. Atom-positional RMSDs were calculated after translational superposition of solute centers of mass and rotational least-squares fitting of the atomic coordinates of all backbone atoms of the alanine deca-peptide. Hydrogen bonds were defined by a maximum H-atom-acceptor distance of 0.25 nm and a minimum donor-H atom-acceptor angle of 135° .

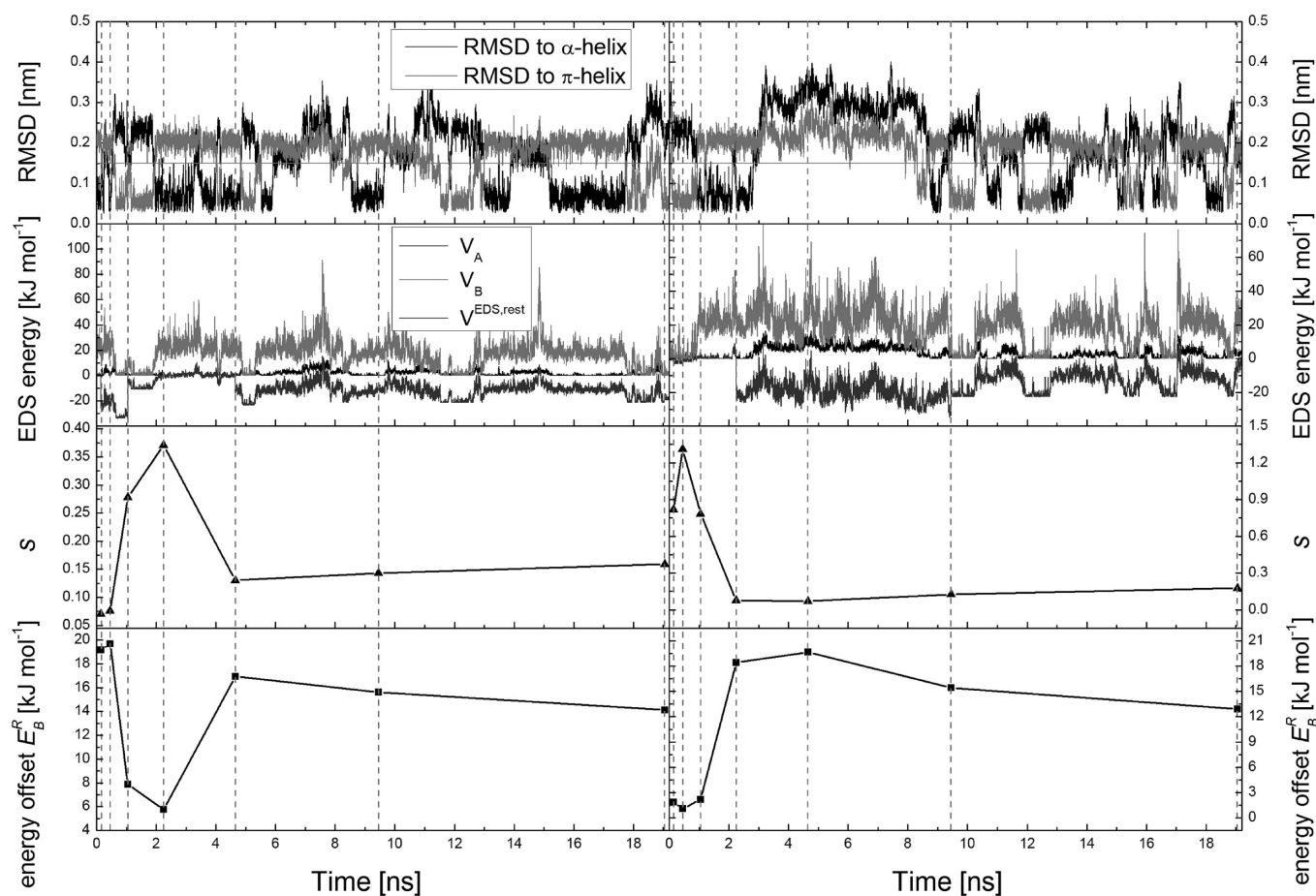


Figure 7. Time evolutions of different properties in the α -helix/ π -helix EDS parameter update simulations *update5* (left panels) and *update6* (right panels), see also Table 1. From top to bottom: backbone atom-positional RMSD of the peptide with respect to the α -helix or π -helix, the restraining and the EDS reference potential energies, smoothness parameter s , energy offset E_B^R . The horizontal gray line represents an RMSD of 0.15 nm used to define set α (α -helix) and set β (π -helix); the vertical dashed lines show when the updates of s and E_B^R are carried out.

4. RESULTS AND DISCUSSION

4.1. Free Enthalpy Estimation from Thermodynamic Integration (TI) Simulations. $\langle(\partial V_{\text{TI}})/(\partial\lambda)\rangle_\lambda$ as a function of λ in the TI simulations is shown in Figure 3a. This curve presents two unusual features. First, there is a sudden change of $\langle(\partial V_{\text{TI}})/(\partial\lambda)\rangle_\lambda$ between λ values 0.4 and 0.5. Second, the first five λ values have a similar value for $\langle(\partial V_{\text{TI}})/(\partial\lambda)\rangle_\lambda$, and the last six λ values have again a similar value. The backbone atom-positional RMSD of the peptide in the simulations at λ values 0.4 and 0.5 are shown in Figure 3c and d together with $(\partial V_{\text{TI}})/(\partial\lambda)$. When $\lambda = 0.4$, the α -helical conformation was mostly sampled, whereas at $\lambda = 0.5$ mainly the π -helical conformation was sampled. This can explain the sharp drop in the $\langle(\partial V_{\text{TI}})/(\partial\lambda)\rangle_\lambda$ curve. Some attempts of transition between the two helices also occurred during the simulations.

If we look closer at the λ -dependent pathway between the two end states of the TI simulation, as shown by a schematic representation of V^{phys} in Figure 3b, state A and state B , which restrain the peptide into the α -helix and the π -helix, respectively, are in different local minima. There is a barrier between them, and the pathway defined by eq 28 crosses the barrier. Thus, at small λ values, the system remained in the local minimum of state A , and mostly the α -helix was sampled. At large λ values, on the other hand, the system stayed in state B , and mostly the π -helix

was sampled. At the intermediate λ values, transitions between two helices should occur but are rare, which affects the convergence of $\langle(\partial V_{\text{TI}})/(\partial\lambda)\rangle_\lambda$. Since the barrier hinders an accurate evaluation of $\langle(\partial V_{\text{TI}})/(\partial\lambda)\rangle_\lambda$ at these λ values, the integral in eq 25 is not very precise.

The free enthalpy difference between two end states B and A , i.e., ΔG_{BA} , obtained by TI using eq 29 is $10.6 \pm 1.1 \text{ kJ mol}^{-1}$. In view of the discussion above, the statistical uncertainty estimated by block averaging⁴⁹ underestimates the real error bar. To calculate the free enthalpy difference between the two conformational sets β (π -helix) and α (α -helix), i.e. $\Delta G_{\pi\alpha}$, reweighting must be applied.

4.2. Free Enthalpy Estimation from End-State Simulations. The distributions of $V_B - V_A$ in the simulations of both end states are shown in Figure 4, in which the crossing point can be used as an estimate for the free enthalpy difference between the two end states B and A , i.e. ΔG_{BA} . The result obtained is 8.1 kJ mol^{-1} . The energy distributions of the first 5 ns and second 5 ns of the simulations differ in the tail parts (inset of Figure 4), which means that the result suffers from statistical uncertainty. In fact, the crossing points of distributions averaged over different simulation periods range from less than 7 kJ mol^{-1} to more than 10 kJ mol^{-1} , indicating that the statistical uncertainty of the result is about 2 kJ mol^{-1} .

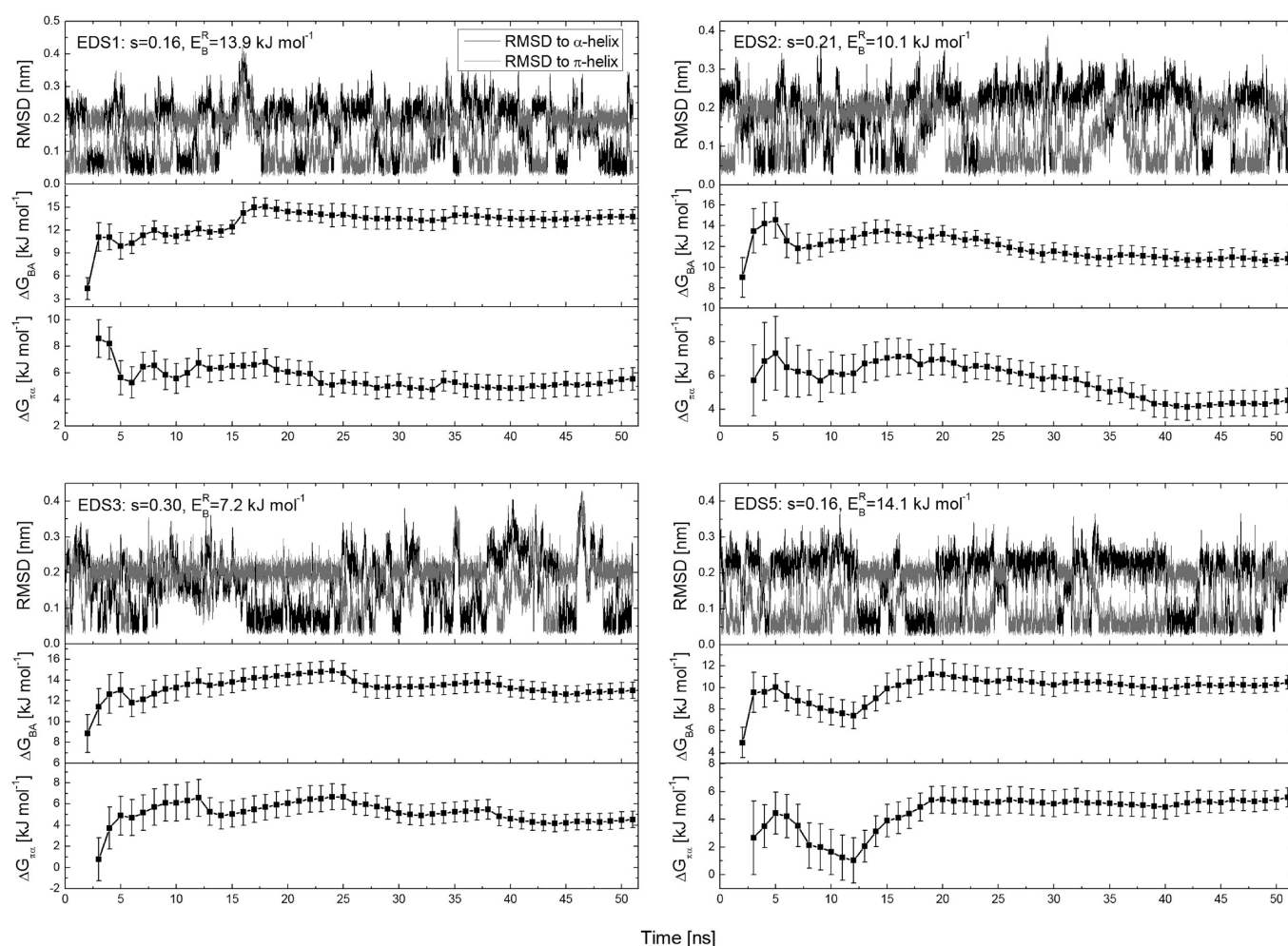


Figure 8. Time evolutions of backbone atom-positional RMSD of the peptide with respect to the α -helix or π -helix, free enthalpy differences of the two end-state Hamiltonians B (π -helix) and A (α -helix), ΔG_{BA} , and free enthalpy differences of the two conformational sets β (π -helix) and α (α -helix), $\Delta G_{\pi\omega}$ in four EDS evaluation simulations.

The free enthalpy differences between the two conformational sets β (π -helix) and α (α -helix) calculated through eq 25 are listed in Table 2 for the four different EDS reference-state Hamiltonians used as intermediate states, of which two have positive s parameters and the other two have negative s parameters. $s > 0$ corresponds to an EDS reference-state Hamiltonian which envelopes the parts of configuration space of both states A and B , while $s < 0$ results in an EDS reference-state Hamiltonian which is a subset of both states A and B .³⁵ The results of EDS reference-state Hamiltonians with the negative s parameters as the intermediate states should be more reliable, because $\langle e^{-(V_M - V_A)/k_B T} \rangle_A$ and $\langle e^{-(V_M - V_B)/k_B T} \rangle_B$ in eq 25 are more accurate if V_M is sampling a subset of the conformations belonging to V_A and V_B , assuming that the two end states have overlapping conformational spaces, and that these are sufficiently sampled. This method is in principle very similar to the overlap sampling method,^{31,32} the only difference being the choice of the intermediate state. Thus, the overlap sampling method can also be generalized to the free energy calculation of different conformational states.

4.3. Free Enthalpy Estimation from Enveloping Distribution Sampling (EDS) Simulations. The backbone atom-positional RMSD of the peptide for the EDS parameter update

simulations *update1* and *update2* are shown in the upper panels of Figure 5. Both α - and π -helices were sampled during the simulations; the typical residence time between transitions was on the time scale of nanoseconds. The smoothness parameter s and the energy offset parameter E_B^R show strong anticorrelations during the update procedure (bottom panels of Figure 5). The evolution of the parameters is highly dependent on the configurations sampled; therefore they show large fluctuations during the update procedure. The same quantities for the EDS parameter update simulations with increasing simulation time periods between updates, *update3* and *update4*, are shown in Figure 6. Longer time periods ensured more transitions between the two helical conformations, thus reducing the fluctuations of the parameters to a certain extent. The exclusion of the less relevant conformations, sets γ and δ , during the update procedure did not influence the fluctuations and therefore did not improve the convergence of the parameters (Figure 7).

The resulting values for the s and E_B^R parameters are 0.16 and 13.9 kJ mol⁻¹, 0.21 and 10.1 kJ mol⁻¹, 0.30 and 7.2 kJ mol⁻¹, 0.29 and 7.4 kJ mol⁻¹, 0.16 and 14.1 kJ mol⁻¹, and 0.18 and 12.9 kJ mol⁻¹ for the six update simulations, respectively. It seems to be difficult to reach convergence to unique optimal parameters, but all of the resulting s and E_B^R parameter values are in a

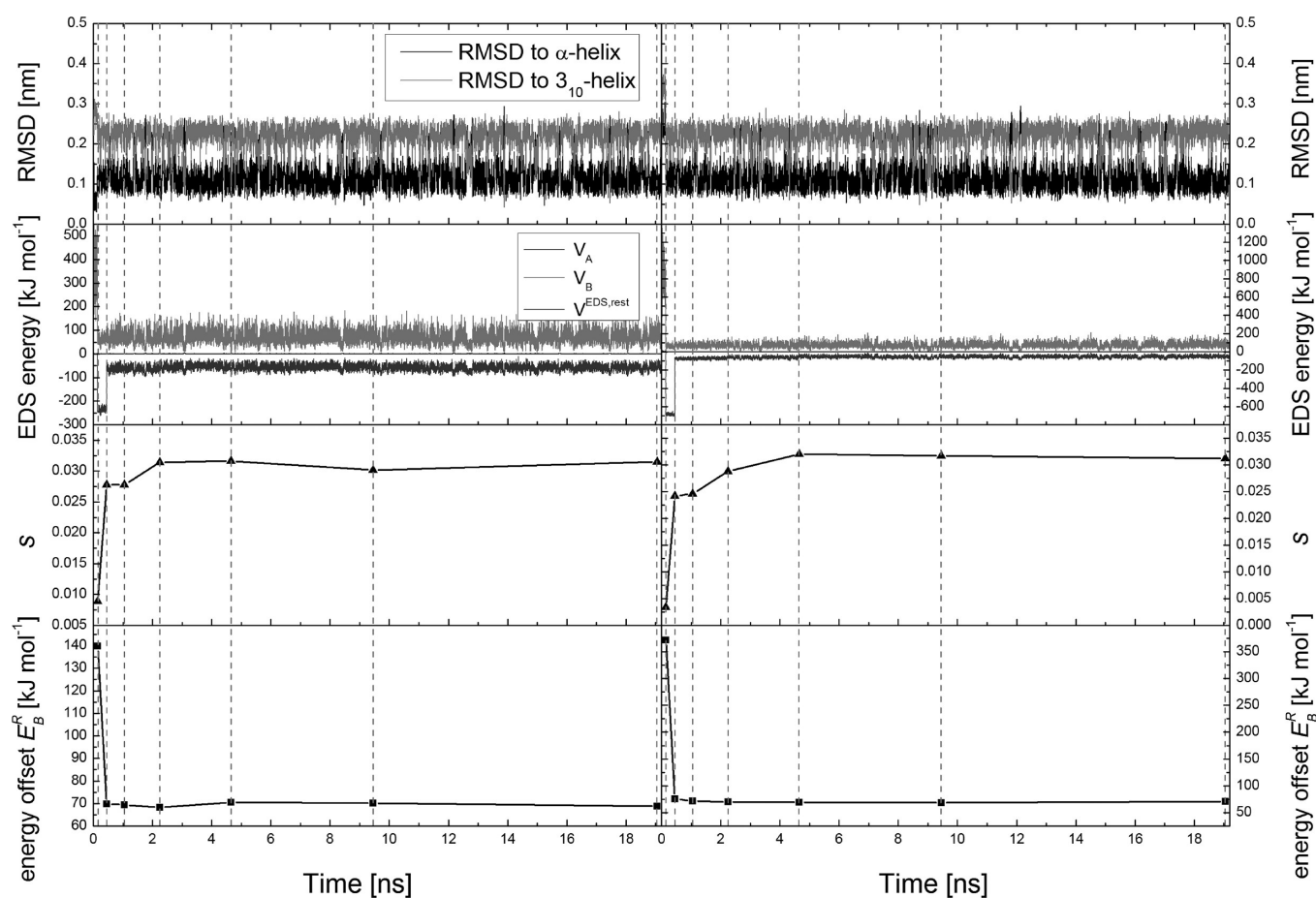


Figure 9. Time evolutions of different properties in the α -helix/ 3_{10} -helix EDS parameter update simulations *update7* (left panels) and *update8* (right panels), see also Table 1. From top to bottom: backbone atom-positional RMSD of the peptide with respect to the α -helix or 3_{10} -helix, the restraining and the EDS reference potential energies, smoothness parameter s , energy offset E_B^R . The vertical dashed lines show when the updates of s and E_B^R are carried out.

reasonable range. Moreover, there may be more than one set of EDS parameters that allows optimal sampling of both end states A and B using the reference-state Hamiltonian H_R . Trying yet other update schemes would not help to converge the parameters, because the limiting factor is the relatively long residence time required at the end states. Longer time periods between updates or the use of enhanced sampling techniques may improve the convergence of the s and E_B^R parameters.

Time evolutions of backbone atom-positional RMSD of the peptide, of ΔG_{BA} and $\Delta G_{\pi\alpha}$ in the four EDS evaluation simulations (Table 1) are shown in Figure 8. ΔG_{BA} was calculated through eq 20 and does not depend explicitly on the number of configurations belonging to each of the conformational sets α and β , whereas $\Delta G_{\pi\alpha}$ was calculated through eq 22 and does depend on them. Nevertheless, the two properties show a similar trend in all four simulations; i.e., the free enthalpy differences tend to increase when set α is sampled and to decrease when set β is sampled. Though having quite different s and E_B^R parameters, the four EDS simulations gave similar pictures, except for the relative populations of the two conformational sets α and β . The curves of ΔG_{BA} become flat after 20–30 ns (Figures 8 and S4, Supporting Information), while the convergence of $\Delta G_{\pi\alpha}$ is slower because the relative population of the conformational set β versus set α (the last term in eq 22) converges slower, i.e., after 20–40 ns

(Figures 8 and S5, Supporting Information). Again, application of enhanced sampling techniques might speed up the convergence of the free enthalpy differences.

The results for ΔG_{BA} and $\Delta G_{\pi\alpha}$ in the four EDS evaluation simulations are shown in Table 2. The difference in ΔG_{BA} among the four simulations is about 3 kJ mol^{-1} . The block averaging used underestimates, with values less than 1 kJ mol^{-1} , the statistical uncertainties of the results. For $\Delta G_{\pi\alpha}$ the difference among the four values is within 1 kJ mol^{-1} . The EDS results for both free enthalpy differences are consistent with the corresponding ones obtained from the TI and end-state simulations.

The results of the EDS calculations to obtain the free enthalpy difference $\Delta G_{310\alpha}$ between a 3_{10} -helix and an α -helix for decalanine in aqueous solution are shown in Figures 9 and 10. Various quantities for two EDS parameter update simulations connecting two end states A (α -helix) and C (3_{10} -helix), *update7* and *update8*, are shown in Figure 9. The convergence of the s and E_B^R parameters is much faster than for the EDS simulations connecting the end states A (α -helix) and B (π -helix), probably due to a larger overlap of the phase spaces between the end states A and C . The resulting s and E_B^R parameters are 0.03 and 68.7 kJ mol^{-1} and 0.03 and 70.5 kJ mol^{-1} for the *update7* and *update8* simulations, respectively. The RMSD, ΔG_{CA} , and $\Delta G_{310\alpha}$ for the EDS evaluation simulation, EDS7, are shown in Figure 10. Although E_B^R is very big and many transition events were

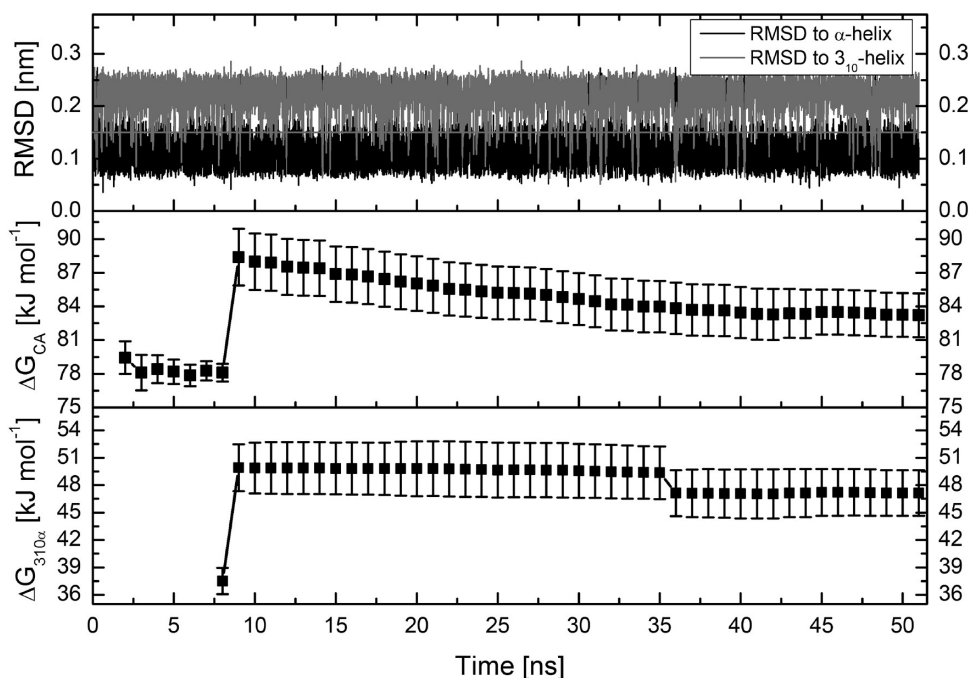


Figure 10. Time evolutions of backbone atom-positional RMSD of the peptide with respect to the α -helix or 3_{10} -helix, free enthalpy differences of the two end-state Hamiltonians C (3_{10} -helix) and A (α -helix), ΔG_{CA} , and free enthalpy differences of the two conformational sets β (3_{10} -helix) and α (α -helix), $\Delta G_{310\alpha}$ in EDS evaluation simulation EDS7.

occurring, the simulation sampled mostly the α -helix. The results are $83.2 \pm 1.9 \text{ kJ mol}^{-1}$ and $47.1 \pm 2.5 \text{ kJ mol}^{-1}$ for ΔG_{CA} and $\Delta G_{310\alpha}$, respectively.

4.4. Free Enthalpy Differences between Three Helices.

According to the EDS simulations, the free enthalpy differences of the π -helix and the 3_{10} -helix versus the α -helix are 5 kJ mol^{-1} and 47 kJ mol^{-1} , respectively, for the alanine deca-peptide using the GROMOS 53A6 force field in SPC water solvent. That is, the α -helix and the π -helix are of similar stability, whereas the 3_{10} -helix is very unstable for deca-alanine. In a different context, studies²¹ have shown that for the CHARMM 22 force field, these values are $-1.9 \text{ kcal mol}^{-1}$ and $20.9 \text{ kcal mol}^{-1}$ for the free enthalpy differences of the π -helix and the 3_{10} -helix versus the α -helix of deca-alanine in TIP3P water solvent. Unfortunately, no corresponding experimental data are available for deca-alanine in aqueous solution.

5. CONCLUSION

In this work, statistical-mechanical expressions were formulated which allow the application of the technique of enveloping distribution sampling (EDS) to the computation of free enthalpy differences between different conformations or conformational states instead of between different Hamiltonians. Different helical conformations, i.e., 3_{10} -, α -, and π -helices, of an alanine deca-peptide in explicit water solvent were considered. The results were compared to those obtained from the standard technique of thermodynamic integration (TI) and from end-state simulations.

A reasonably accurate result was obtained using TI simulations, in which the λ -dependent pathway was defined on the basis of the hydrogen bonds characterizing the different helical conformations. Because this pathway involves an energetic barrier in the physical potential energy V^{phys} , a large change occurred in the

$\langle(\partial V_{\text{TI}})/(\partial \lambda)\rangle_{\lambda}$ curve for λ values at the barrier. Moreover, it is difficult to converge $\langle(\partial V_{\text{TI}})/(\partial \lambda)\rangle_{\lambda}$ in the simulations at intermediate λ values due to the instability of the conformations near the barrier. Thus, the TI result suffers from statistical uncertainty. One may argue that TI might have produced a better result if a better pathway, i.e., λ -dependence of the biasing potential energy terms, had been defined. However, often a “better” pathway that avoids a barrier is not so easily found. This is one of the drawbacks of the TI method.

This situation can be avoided by estimating the free enthalpy difference based on end-state simulations. However, in that case, overlap of conformational space between the two end states is required for such calculations to yield meaningful results.

In the two-state EDS simulations, an unphysical reference-state Hamiltonian which allows sampling of the conformations relevant to the different end states is constructed automatically. Both helices were sampled during the simulations, and the transitions between them were on the time scale of nanoseconds, which makes the convergence of the EDS smoothness parameter and the EDS energy offset parameter to their “optimal” values slow. However, the resulting EDS parameters of the different EDS parameter update simulations were in a reasonable range. Extending the time period between updates did speed up the convergence, whereas the exclusion of conformations not relevant to the two conformations did not improve the convergence in this case.

The EDS evaluation simulations using these EDS reference states were used to calculate the free enthalpy differences between the two end states and between the two helical conformations. The different EDS reference-state Hamiltonians gave similar results. According to the EDS simulations, for the deca-alanine using the GROMOS 53A6 force field and SPC water as the solvent, the α -helix and the π -helix are of similarly stability, whereas the 3_{10} -helix is very unstable.

The presented EDS method can be considered as a particular type of umbrella sampling, with an automatically generated biasing umbrella potential energy function of the particular form eq 2, which is solely based on end-state potential energy functions and which ensures the sampling of both end-state conformations. We have shown that the EDS method is applicable to computing conformational free energy differences as well as alchemical ones. Only two-state problems were considered, but the method can be generalized to multiple-state problems and can be further optimized by combination with techniques to speed up the sampling of conformational space.

APPENDIX

Equation 22 can be derived as follows. According to eq 21, we have

$$\Delta G_{\beta\alpha} = G_{\beta} - G_{\alpha}$$

$$= -k_B T \ln \left\{ \frac{\int e^{-V^{\text{phys}}(\bar{r}^N)/k_B T} \delta(\bar{r}^N - \bar{r}_{\beta}^N) d\bar{r}^N}{\int e^{-V^{\text{phys}}(\bar{r}^N)/k_B T} \delta(\bar{r}^N - \bar{r}_{\alpha}^N) d\bar{r}^N} \right\} \quad (31)$$

where $\delta(\bar{r}^N - \bar{r}_{\xi}^N)$ is the delta function that selects configurations that belong to set ξ . Using eq 2, we can rewrite eq 31 as

$$\Delta G_{\beta\alpha} = -k_B T$$

$$\times \ln \left\{ \frac{\int e^{+V^{\text{EDS,rest}}(\bar{r}^N)/k_B T} \times e^{-V_R(\bar{r}^N)/k_B T} \times \delta(\bar{r}^N - \bar{r}_{\beta}^N) d\bar{r}^N}{\int e^{-V_R(\bar{r}^N)/k_B T} \times \delta(\bar{r}^N - \bar{r}_{\beta}^N) d\bar{r}^N} \right.$$

$$\times \frac{\int \delta(\bar{r}^N - \bar{r}_{\beta}^N) \times e^{-V_R(\bar{r}^N)/k_B T} d\bar{r}^N}{\int e^{-V_R(\bar{r}^N)/k_B T} d\bar{r}^N}$$

$$\times \frac{\int e^{-V_R(\bar{r}^N)/k_B T} d\bar{r}^N}{\int \delta(\bar{r}^N - \bar{r}_{\alpha}^N) \times e^{-V_R(\bar{r}^N)/k_B T} d\bar{r}^N}$$

$$\left. \times \frac{\int \delta(\bar{r}^N - \bar{r}_{\alpha}^N) \times e^{-V_R(\bar{r}^N)/k_B T} d\bar{r}^N}{\int e^{+V^{\text{EDS,rest}}(\bar{r}^N)/k_B T} \times e^{-V_R(\bar{r}^N)/k_B T} \times \delta(\bar{r}^N - \bar{r}_{\alpha}^N) d\bar{r}^N} \right\}$$

$$= -k_B T \ln \left\{ \frac{\langle e^{+V^{\text{EDS,rest}}/k_B T} \rangle_{R, \text{set}\beta}}{\langle e^{+V^{\text{EDS,rest}}/k_B T} \rangle_{R, \text{set}\alpha}} \times \frac{\langle \delta(\bar{r}^N - \bar{r}_{\beta}^N) \rangle_R}{\langle \delta(\bar{r}^N - \bar{r}_{\alpha}^N) \rangle_R} \right\} \quad (32)$$

Using eq 23 in eq 32, we find eq 22.

Using the definition $V_X = V^{\text{phys}} + V_X^{\text{rest}}$ for $X = A$ and $X = B$, eq 31 can also be rewritten as

$$\Delta G_{\beta\alpha} = -k_B T$$

$$\times \ln \left\{ \frac{\int e^{+V_B^{\text{rest}}(\bar{r}^N)/k_B T} \times e^{-V_B(\bar{r}^N)/k_B T} \times \delta(\bar{r}^N - \bar{r}_{\beta}^N) d\bar{r}^N}{\int e^{-V_B(\bar{r}^N)/k_B T} \times \delta(\bar{r}^N - \bar{r}_{\beta}^N) d\bar{r}^N} \right.$$

$$\times \frac{\int \delta(\bar{r}^N - \bar{r}_{\beta}^N) \times e^{-V_B(\bar{r}^N)/k_B T} d\bar{r}^N}{\int e^{-V_B(\bar{r}^N)/k_B T} d\bar{r}^N}$$

$$\left. \times \frac{\int e^{-V_B(\bar{r}^N)/k_B T} d\bar{r}^N}{\int e^{+V_A^{\text{rest}}(\bar{r}^N)/k_B T} \times e^{-V_A(\bar{r}^N)/k_B T} \times \delta(\bar{r}^N - \bar{r}_{\alpha}^N) d\bar{r}^N} \right\}$$

$$\times \frac{\int e^{-V_B(\bar{r}^N)/k_B T} d\bar{r}^N}{\int e^{-(V_M(\bar{r}^N) - V_B(\bar{r}^N))/k_B T} \times e^{-V_B(\bar{r}^N)/k_B T} d\bar{r}^N}$$

$$\times \frac{\int e^{-(V_M(\bar{r}^N) - V_A(\bar{r}^N))/k_B T} \times e^{-V_A(\bar{r}^N)/k_B T} d\bar{r}^N}{\int e^{-V_A(\bar{r}^N)/k_B T} d\bar{r}^N}$$

$$\times \frac{\int e^{-V_A(\bar{r}^N)/k_B T} d\bar{r}^N}{\int \delta(\bar{r}^N - \bar{r}_{\alpha}^N) \times e^{-V_A(\bar{r}^N)/k_B T} d\bar{r}^N}$$

$$\left. \times \frac{\int e^{-V_A(\bar{r}^N)/k_B T} \times \delta(\bar{r}^N - \bar{r}_{\alpha}^N) d\bar{r}^N}{\int e^{+V_A^{\text{rest}}(\bar{r}^N)/k_B T} \times e^{-V_A(\bar{r}^N)/k_B T} \times \delta(\bar{r}^N - \bar{r}_{\alpha}^N) d\bar{r}^N} \right\}$$

$$= -k_B T \ln \left\{ \frac{\langle e^{+V_B^{\text{rest}}/k_B T} \rangle_{B, \text{set}\beta}}{\langle e^{+V_A^{\text{rest}}/k_B T} \rangle_{A, \text{set}\alpha}} \times \frac{\langle 1 \rangle_{B, \text{set}\beta}}{\langle 1 \rangle_{A, \text{set}\alpha}} \times \frac{\langle e^{-(V_M - V_A)/k_B T} \rangle_A}{\langle e^{-(V_M - V_B)/k_B T} \rangle_B} \right\} \quad (33)$$

which is eq 25.

ASSOCIATED CONTENT

S Supporting Information. The RMSD curves and the hydrogen-bond populations of the end-state simulations are available. This information is available free of charge via the Internet at <http://pubs.acs.org/>.

AUTHOR INFORMATION

Corresponding Author

*E-mail: wfvgn@igc.phys.chem.ethz.ch.

ACKNOWLEDGMENT

This work was financially supported by the National Center of Competence in Research (NCCR) in Structural Biology and by grant number 200020-121913 of the Swiss National Science Foundation, by grant number 228076 of the European Research Council (ERC), and by grants numbers IZLCZ2-123884 (Swiss) and GJHZ0906 (Chinese) of the Sino–Swiss Science and Technology Cooperation Program, which is gratefully acknowledged.

REFERENCES

- (1) Straatsma, T. P.; McCammon, J. A. *Annu. Rev. Phys. Chem.* **1992**, *43*, 407–435.
- (2) Kollman, P. *Chem. Rev.* **1993**, *93*, 2395–2417.
- (3) Simonson, T.; Archontis, G.; Karplus, M. *Acc. Chem. Res.* **2002**, *35*, 430–437.
- (4) van Gunsteren, W. F.; Daura, X.; Mark, A. E. *Helv. Chim. Acta* **2002**, *85*, 3113–3129.
- (5) Kofke, D. A. *Fluid Phase Equilib.* **2005**, *228*, 41–48.
- (6) Rodinger, T.; Pomes, R. *Curr. Opin. Struct. Biol.* **2005**, *15*, 164–170.
- (7) Gilson, M. K.; Zhou, H. X. *Annu. Rev. Biophys. Biomol. Struct.* **2007**, *36*, 21–42.
- (8) Meirovitch, H. *Curr. Opin. Struct. Biol.* **2007**, *17*, 181–186.
- (9) Jorgensen, W. L.; Thomas, L. L. *J. Chem. Theory Comput.* **2008**, *4*, 869–876.
- (10) Knight, J. L.; Brooks, C. L. *J. Comput. Chem.* **2009**, *30*, 1692–1700.

- (11) Vanden-Eijnden, E. *J. Comput. Chem.* **2009**, *30*, 1737–1747.
- (12) Christ, C. D.; Mark, A. E.; van Gunsteren, W. F. *J. Comput. Chem.* **2010**, *31*, 1569–1582.
- (13) Postma, J. P. M.; Berendsen, H. J. C.; Haak, J. R. *Faraday Symp. Chem. Soc.* **1982**, *17*, 55–67.
- (14) Grossfield, A.; Ren, P. Y.; Ponder, J. W. *J. Am. Chem. Soc.* **2003**, *125*, 15671–15682.
- (15) Jorgensen, W. L. *Science* **2004**, *303*, 1813–1818.
- (16) Oostenbrink, C.; van Gunsteren, W. F. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6750–6754.
- (17) Daura, X.; Jaun, B.; Seebach, D.; van Gunsteren, W. F.; Mark, A. E. *J. Mol. Biol.* **1998**, *280*, 925–932.
- (18) Brooks, C. L. *Acc. Chem. Res.* **2002**, *35*, 447–454.
- (19) Roux, B.; Karplus, M. *Biophys. J.* **1991**, *59*, 961–981.
- (20) Hansen, H. S.; Hunenberger, P. H. *J. Chem. Theory Comput.* **2010**, *6*, 2622–2646.
- (21) Mahadevan, J.; Lee, K. H.; Kuczera, K. *J. Phys. Chem. B* **2001**, *105*, 1863–1876.
- (22) Wang, J.; Gu, Y.; Liu, H. Y. *J. Chem. Phys.* **2006**, *125*, 094907.
- (23) Torrie, G. M.; Valleau, J. P. *J. Comput. Phys.* **1977**, *23*, 187–199.
- (24) Hansen, H. S.; Hunenberger, P. H. *J. Comput. Chem.* **2010**, *31*, 1–23.
- (25) Kirkwood, J. G. *J. Chem. Phys.* **1935**, *3*, 300–313.
- (26) Zwanzig, R. W. *J. Chem. Phys.* **1954**, *22*, 1420–1426.
- (27) Powles, J. G.; Evans, W. A. B.; Quirke, N. *Mol. Phys.* **1982**, *46*, 1347–1370.
- (28) Shing, K. S.; Gubbins, K. E. *Mol. Phys.* **1982**, *46*, 1109–1128.
- (29) Bennett, C. H. *J. Comput. Phys.* **1976**, *22*, 245–268.
- (30) Konig, G.; Bruckner, S.; Boresch, S. *J. Comput. Chem.* **2009**, *30*, 1712–1718.
- (31) Lu, N. D.; Kofke, D. A.; Woolf, T. B. *J. Comput. Chem.* **2004**, *25*, 28–39.
- (32) Lu, N. D.; Singh, J. K.; Kofke, D. A. *J. Chem. Phys.* **2003**, *118*, 2977–2984.
- (33) Liu, H. Y.; Mark, A. E.; van Gunsteren, W. F. *J. Phys. Chem.* **1996**, *100*, 9485–9494.
- (34) Christ, C. D.; van Gunsteren, W. F. *J. Chem. Phys.* **2007**, *126*, 184110.
- (35) Christ, C. D.; van Gunsteren, W. F. *J. Chem. Phys.* **2008**, *128*, 174112.
- (36) Christ, C. D.; van Gunsteren, W. F. *J. Chem. Theory Comput.* **2009**, *5*, 276–286.
- (37) Christ, C. D.; van Gunsteren, W. F. *J. Comput. Chem.* **2009**, *30*, 1664–1679.
- (38) Riniker, S.; Christ, C. D.; Hansen, N.; Mark, A. E.; Nair, P. C.; van Gunsteren, W. F. *J. Chem. Phys.* **2011**, *135*, 024105.
- (39) Hansen, N.; Dolenc, J.; Knecht, M.; Riniker, S.; van Gunsteren, W. F. *J. Comput. Chem.* **2011**, accepted.
- (40) Lin, Z. X.; Kornfeld, J.; Machler, M.; van Gunsteren, W. F. *J. Am. Chem. Soc.* **2010**, *132*, 7276–7278.
- (41) Oostenbrink, C.; Villa, A.; Mark, A. E.; van Gunsteren, W. F. *J. Comput. Chem.* **2004**, *25*, 1656–1676.
- (42) 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; pp 331–342.
- (43) van Gunsteren, W. F.; Billeter, S. R.; Eising, A. A.; Hunenberger, P. H.; Krüger, P.; Mark, A. E.; Scott, W. R. P.; Tironi, I. G. *Biomolecular Simulation: The GROMOS96 Manual and User Guide*; Vdf Hochschulverlag AG an der ETH Zürich: Zürich, Switzerland, 1996.
- (44) Christen, M.; Hunenberger, 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.
- (45) Riniker, S.; Christ, C. D.; Hansen, H. S.; Hunenberger, P. H.; Oostenbrink, C.; Steiner, D.; van Gunsteren, W. F. *J. Phys. Chem. B* **2011**; DOI: 10.1021/jp204303a.
- (46) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684–3690.
- (47) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *23*, 327–341.
- (48) Tironi, I. G.; Sperb, R.; Smith, P. E.; van Gunsteren, W. F. *J. Chem. Phys.* **1995**, *102*, 5451–5459.
- (49) Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids*; Oxford University Press: New York, 1987; pp 191–195.

Accuracy of Density Functionals in the Prediction of Electronic Proton Affinities of Amino Acid Side Chains

Natércia F. Brás,^{†,§} Marta A. S. Perez,^{†,§} Pedro A. Fernandes,[†] Pedro J. Silva,[‡] and Maria J. Ramos^{*,†}

[†]REQUIMTE, Departamento de Química, Faculdade de Ciências, Universidade do Porto, Rua do Campo Alegre, 687, 4169-007 Porto, Portugal

[‡]REQUIMTE, Fac. de Ciências da Saúde, Universidade Fernando Pessoa, Rua Carlos da Maia, 296, 4200-150 Porto, Portugal

 Supporting Information

ABSTRACT: The ionization states of amino acids influence the structure, function, stability, solubility, and reactivity of proteins and are difficult to determine unambiguously by experimental means. Thus, it is very important to be able to determine them theoretically and with high reliability. We have analyzed how well DFT functionals, often used to characterize complex and large models such as proteins, describe the zero-point-exclusive proton affinity at 0 K, PA_{el}^{OK} , for the ionizable side chains of lysine (Lys), histidine (His), arginine (Arg), and aspartate (Asp^-) as well as the cysteine (Cys^-), serine (Ser^-), and tyrosine (Tyr^-) anions. The reference values PA_{el}^{OK} were determined at the very accurate CCSD(T)/CBS level. Those values were obtained by the sum of the complete basis set limit of the MP2 energies plus a CCSD(T) correction term evaluated with the aug-cc-pVTZ basis set. The complete basis set limit of MP2 energies was determined using the Truhlar and Helgaker extrapolation schemes. A new, important, and consistent DFT benchmarking database for PA_{el}^{OK} and for proton transfer between two different ionizable side chains, ΔPA_{el}^{OK} , is provided, making this work relevant to all studies with ionizable amino acids side chains that use DFT. Among the 64 density functionals tested, the MPW1B95-D3, XYG3, MPW1B95, B1B95-D3, BMK, BMK-D3, M06-2X, B1LYP, B1B95, PBE1PBE, CAM-B3LYP, B97-1, PBE1KCIS, B3P86, CAM-B3LYP-D3, B3LYP, B98, M06-L, and M06 provide the most accurate PA_{el}^{OK} values for all ionizable amino acids studied, with errors below 1.5 kcal/mol, which translate into an error of less than 1 pK_a unit in solution. Furthermore, among the best rated to predict PA_{el}^{OK} , we have found that M06-2X was the most accurate density functional for proton transfers between different amino acids.

INTRODUCTION

Amino acids are critical to life. They are the building units of proteins and have countless fundamental functions in the metabolism of all living beings. The amino acid residues with ionizable side chains make up, on average, 29% of the residues in proteins.¹ Their ionization state influences the structure, function, stability, solubility, and reactivity of proteins.^{2,3} Solvent exposure, Bohr effect, charge–dipole interactions, and charge–charge interactions are important environmental effects that change the ionization state of the amino acid residues in proteins⁴ and, consequently, the charge of the proteins. The enzymatic reactivity strongly depends on the ionization state of the active site residues and on the proton transfers taking place between them. The recognition of ligands and drugs by biological receptors is strongly influenced by the charge of the receptor. The knowledge of the ionization state of amino acid residues in proteins is thus of great importance to the chemical and pharmaceutical industries.

Experimental methods that assign the ionization state to the amino acid residues are typically indirect and difficult to perform.^{5,6} While the proton affinities of a large variety of organic molecules have been measured, the data on amino acid residues in proteins is less complete, in part because they tend to be relatively nonvolatile and thermally labile.⁵ Moreover, the methods of measuring proton affinities (equilibrium measurements, reaction bracketing, and kinetic methods) provide only relative

values. As there is still significant uncertainty in the reference scale of proton affinities, it becomes difficult to rank them on an absolute scale.⁵ The proton affinity of a given amino acid can vary significantly in different protein environments, which makes difficult the unambiguous determination of its ionization state. The fact that the protons are not detectable by X-ray crystallography further complicates this issue.^{5,6} As experimental methods face severe impracticalities in many situations it becomes very important to calculate the protonation state of the ionizable amino acid side chains.

There are numerous studies that report different kinds of theoretical approaches and procedures, which predict the protonation state of ionizable residues.^{7–12} Most of them predict the pK_a and/or the proton affinities (PA) and/or the gas-phase basicities (GPB) in a specific environment and compare the predictions with experimental values. Many of these methods are knowledge-based.^{11,12} Even though this philosophy is quite useful, knowledge-based methods always have the limitation of depending on the training set that is used for the calibration and typically fail in the most complex and unusual cases, which constitute the most interesting ones from a biochemistry point of view. Moreover, given the extraordinary diversity of micro-environments that proteins can build up, knowledge-based

Received: May 4, 2011

Published: October 25, 2011

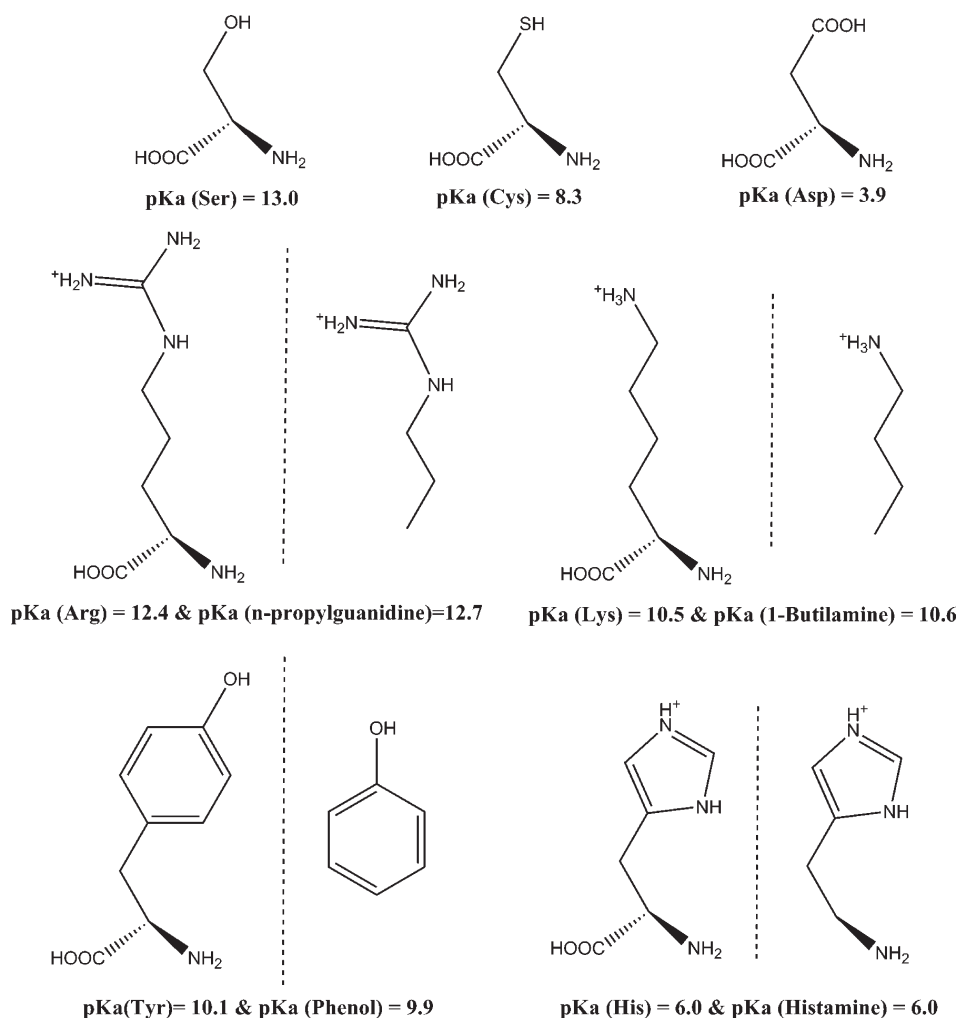


Figure 1. Representation of the structures and pK_a values of all ionizable amino acids and corresponding models.

approaches can never be as robust as would be desirable in this context. This said, we admit that in many instances, within the actual level of development of electronic structure calculations, knowledge-based methods can lead to smaller average errors, as the calculation of long-range interactions and the inclusion of flexibility in proteins is prohibitively time-consuming with electronic structure methods. However, electronic structure calculations have the potential of retrieving extremely precise results, both in usual and unusual environments with comparable accuracy. This is the basic reason why the scientific community is investing so much effort to predict protonation states with high-level electronic structure calculations. Although some studies have been made with wave function methods,^{7,13,14} the most frequently used methods to determine the protonation state of the residues are based on Density Functional Theory (DFT),^{10,15} which can be quite effective with much less computational time if the proper functionals are chosen.

However, it is not trouble-free to properly choose a DFT functional among the very large number currently available. Moreover, new density functionals continue to appear, taking the performance of DFT to new levels of accuracy. The selection of the functional to use will have to depend on the problem at hand. To correctly choose the DFT functionals, we should look for new benchmarking studies that test several density

functionals against experimental data or higher-level methods on the specific system under consideration.^{15–20} Our goal here is to analyze how well DFT functionals describe the protonation state of the ionizable side chains of the amino acids Lys, His, Arg, Asp, Cys, Ser, and Tyr.

The proton affinity at a given temperature is, by definition, the enthalpy of reactions 1 or 2, depending on whether the base is negatively charged or not.



In this work, instead of calculating the proton affinity (PA), we have calculated the zero-point-exclusive proton affinity at 0 K (the electronic proton affinity, $\text{PA}_{\text{el}}^{\text{OK}}$) for the ionizable amino acid side chains. This corresponds to the difference in the electronic energy between the deprotonated and protonated side chains only.

We have not calculated the entropy of the process (which would be needed in order to calculate gas-phase acidities/basicities) because its translational and rotational contributions are equivalent for all of the functionals, and the zero point energy

(ZPE) (and consequently the vibrational entropy) is not the quantity that we intend to benchmark. Additionally, by using the same geometry for all functionals, we can compare more clearly the results with high-level CCSD(T) calculations, without contributions from different geometries or different vibrational frequencies.

The difference in the electronic energy between the deprotonated and protonated side chains is by far the dominant and the most sensitive factor in the proton affinity. It is also the quantity that has the slower convergence along with the quality of the theoretical level.

We have used two *ab initio* methods, MP2 (second Moller–Plesset perturbation theory) and CCSD(T) (single and double excitation coupled cluster theory with perturbative triple correction), and extrapolated the energy to the complete basis set limit (CBS) to compute the PA_{el}^{OK} at the very accurate CCSD(T)/CBS level.

The comparison of the extrapolated CCSD(T)/CBS PA_{el}^{OK} with the results obtained from the calculations with 64 density functionals provides the capabilities and limitations of current DFT functionals to describe the ionizable amino acid side chains. Therefore, we deliver an important and consistent DFT benchmarking database, making this work relevant to all studies with ionizable amino acids side chains in which the use of very accurate post-HF methods is not yet feasible, e.g., most mechanistic studies with biological systems.

COMPUTATIONAL DETAILS

Models of the Ionizable Amino Acids Side Chains. We have used the complete molecules of the smallest amino acids Asp, Ser, and Cys, while for the bigger Tyr, Arg, Lys, and His (for which the use of the CCSD(T) methods with very large basis sets is impractical), we have used models that represent the ionizable amino acid side chains. All molecules have been modeled with neutral terminations to mimic better their electronic structure when they are inserted in a protein backbone. Figure 1 shows the molecules used here. The models that represent the Tyr, Lys, Arg, and His side chains have an experimental pK_a very close to the respective amino acid, therefore constituting excellent models to study the pK_a of the side chains of the real residues. A phenol molecule ($pK_a = 9.9$)²¹ models the Tyr side chain ($pK_a = 10.1$).²² An *n*-propylguanidine ($pK_a = 12.7$) models the Arg side chain ($pK_a = 12.4$),²³ and a 1-butylamine ($pK_a = 10.6$)²⁴ models the Lys side chain ($pK_a = 10.5$).²⁵ Concerning His, we have chosen a histamine molecule as a model because both have the same pK_a value ($pK_a = 6.0$).²⁶ The side chain of this amino acid has one δ nitrogen atom (near the backbone) and one ϵ nitrogen atom (far from backbone), which have a similar tendency to lose their protons (the difference in energy between the two protonated species is approximately 0.5 kcal/mol).⁷ Therefore, we have studied only the deprotonation of the ϵ nitrogen atom. This article is focused on these structures/models, and therefore whenever the amino acids are mentioned throughout the text, these models are meant. Figure 1 shows all of the molecules used in the calculations.

Post-HF *ab Initio* Calculations. We have started by fully optimizing the geometry of each amino acid in implicit solvent at the MP2/6-311+G(d,p) level. It would have been more consistent to calculate the proton affinities with the minimum-energy geometry given by each of the density functionals instead of using MP2 geometries. However, the optimization at the MP2/6-311

+G(d,p) theoretical level is known to give good geometries, and it is not biased toward any functional. The solvent was modeled through the Polarizable Continuum Model (PCM) using the polarizable conductor calculation model (CPCM)^{27,28} and the dielectric constant and radius of water. The purpose was to avoid artificial intramolecular interactions due to the lack of environment. The optimized geometries are shown in the Supporting Information. There are no internal hydrogen bonds within the molecules.

For each optimized structure, we have carried out single point energy calculations in a vacuum at the frozen-core CCSD(T) and frozen-core MP2 levels using the Pople basis sets 6-31G*(0.25), 6-31+G(d), and 6-31+G(d,p) and the Dunning basis sets aug-cc-pVDZ and aug-cc-pVTZ. The modified 6-31G*(0.25) basis set includes d-polarization functions with an exponent of 0.25, which makes them more diffuse than the standard d orbitals and simultaneously accounts for polarization and diffusion. We have also calculated the energies with the aug-cc-pVQZ basis set at the MP2 level. Considering the large size of our models, it was unfeasible to use the aug-cc-pVQZ basis set with CCSD(T). The aug-cc-pVTZ basis set may give less accurate results for the sulfur atom present in the Cys side chain.²⁹ However, we will only analyze energy differences and not the absolute energies. Therefore, most of these eventual inaccuracies will cancel out when we calculate PA_{el}^{OK} .

Extrapolation of the Energies to the Basis Set Limit. The Dunning basis sets constitute hierarchical sequences that enable us to extrapolate the correlation energy to the CBS limit.

To extrapolate the energies to the MP2/CBS level, we have employed both the Truhlar and the Helgaker extrapolation schemes. The total energy at the CBS limit ($E_{MP2/CBS}$) was obtained as the sum of the CBS Hartree–Fock energy ($E_{HF/CBS}$) and the CBS correlation energy ($E_{corr/CBS}$) (eq 3).

$$E_{MP2/CBS} = E_{HF/CBS} + E_{corr/CBS} \quad (3)$$

The extrapolations of the Hartree–Fock energy and of the correlation energy were made independently because the convergence characteristics are very different, with the Hartree–Fock energy converging much faster than the correlation energy. The Truhlar extrapolation employed here uses the aug-cc-pVXZ (X = 2 and 3) basis sets (see eq 4).^{30,31} The parameters α and β in eq 4 have the values of 4.93 and 2.13.³² $E_{corr/DZ}$ energies were obtained with MP2/aug-cc-pVDZ, and $E_{corr/TZ}$ energies were obtained with MP2/aug-cc-pVTZ.

$$E_{MP2/CBS}^{Truhlar} = \frac{3^\alpha}{3^\alpha - 2^\alpha} E_{HF/TZ} - \frac{2^\alpha}{3^\alpha - 2^\alpha} E_{HF/DZ} + \frac{3^\beta}{3^\beta - 2^\beta} E_{corr/TZ} - \frac{2^\beta}{3^\beta - 2^\beta} E_{corr/DZ} \quad (4)$$

While the HF energy converges rather quickly with the basis set size, the slow basis set convergence of the correlation energy is usually considered to be the major obstacle in obtaining CBS energies.^{31–35} However, even though the correlation energy converges slower, it seems to do so in a more systematic way, and thus it may give rise to a lower extrapolation error.

To extrapolate the correlation energy to the CBS limit using larger basis sets (beyond the triple- ζ quality), we have also used the Helgaker et al. extrapolation scheme.³⁶ The corresponding equation when the aug-cc-pVTZ and aug-cc-pVQZ basis sets are used is presented in eq 5, in which $E_{corr/TZ}$ is the correlation

energy at the MP2/aug-cc-pVTZ level and $E_{\text{corr}/\text{QZ}}$ is the correlation energy at the MP2/aug-cc-pVQZ level.

$$E_{\text{corr}/\text{CBS}} = \frac{4^3 E_{\text{corr}/\text{QZ}} - 3^3 E_{\text{corr}/\text{TZ}}}{4^3 - 3^3} \quad (5)$$

We could have used the correlation energies resulting from the aug-cc-pVDZ and aug-cc-pVTZ basis sets in this extrapolation. However, this would lead to very large errors in the extrapolated results.³⁷

It is generally considered that the extrapolation error is dominated by the correlation energy. This tendency is also clearly present in our study (please see the Supporting Information). Therefore, the choice of the method to extrapolate the HF energy (typically polynomial or exponential) does not strongly affect the results. Here, we have used the Truhlar scheme to extrapolate the HF energy in all cases. We also tested the method of Helgaker.³⁸ However, its coefficient was optimized with results calculated using the cc-pVXZ basis sets, and therefore it was less appropriate for our calculations than the method of Truhlar. Nevertheless, the difference between the HF/CBS values using both extrapolation schemes was small (mean unsigned difference of 0.10 kcal/mol). Probably a significant part of the difference does not come from the extrapolation schemes themselves but from the use of a coefficient optimized for the cc-pVXZ basis sets.

The final CBS energy corresponds to the sum of the CBS HF energy, extrapolated using the Truhlar scheme, and the CBS correlation energy, extrapolated with the Helgaker et al.:

$$E_{\text{MP2}/\text{CBS}}^{\text{Helgaker}} = E_{\text{HF}/\text{CBS}} + E_{\text{corr}/\text{CBS}}^{\text{Helgaker}} \quad (6)$$

We have calculated the MP2/CBS correlation energies with both extrapolation schemes discussed above, for all amino acid side chains considered in this work.

The difference between the CCSD(T) and MP2 energies ($\Delta E_{\text{CCSD(T)}-\text{MP2}}$) has a small basis set dependence in many chemical systems (see, e.g., reference³⁹). To estimate the CCSD(T)/CBS energy, we have calculated $\Delta E_{\text{CCSD(T)}-\text{MP2}}$ and added this difference to the MP2/CBS energy. This is a standard procedure, based on the assumption that the difference between MP2 and CCSD(T) has a small dependence on the basis set. The CCSD(T)/CBS electronic energy values can thus be given by eq 7.

$$E_{\text{CCSD(T)}/\text{CBS}} = E_{\text{MP2}/\text{CBS}}^{\text{Helgaker}} + \Delta E_{(\text{CCSD(T)}-\text{MP2})/\text{TZ}} \quad (7)$$

To evaluate how dependent from the basis set this term is, we have calculated $\Delta E_{\text{CCSD(T)}-\text{MP2}}$ with the 6-31G*(0.25), 6-31+G(d), and 6-31+G(d,p) and the aug-cc-pVDZ and aug-cc-pVTZ basis sets.

We could have calculated the CCSD(T)/CBS energy extrapolating directly from the CCSD(T)/aug-cc-pVDZ and aug-cc-pVTZ values (without using the MP2 energies), using the method of Helgaker et al.³⁷ However, the procedure was shown to lead to significant errors, due to the use of the CCSD(T)/aug-cc-pVDZ energy in the extrapolation, which is still too far from the CBS limit to be beneficial for the extrapolation.^{37,39} We tested this method, and indeed we could find differences of about 2 kcal/mol in relation to the more reliable extrapolation based on eq 7. To benchmark the density functionals, we have used $\Delta E_{\text{CCSD(T)}-\text{MP2}}$ calculated with the aug-cc-pVTZ basis set. The other basis sets were used only for the purpose of studying

the convergence of $\Delta E_{\text{CCSD(T)}-\text{MP2}}$. Therefore, whenever we refer simply to the CCSD(T)/CBS energy in this work, we are assuming that $\Delta E_{\text{CCSD(T)}-\text{MP2}}$ was calculated with the largest basis set. The reference value of $\text{PA}_{\text{el}}^{\text{OK}}$ for the DFT benchmarking is defined as eq 8:

$$\text{PA}_{\text{el}}^{\text{OK}} = E_{\text{CCSD(T)}/\text{CBS}}^{\text{A}^-} - E_{\text{CCSD(T)}/\text{CBS}}^{\text{HA}} \quad (8)$$

The differences in $\text{PA}_{\text{el}}^{\text{OK}}$ between amino acid pairs ($\Delta \text{PA}_{\text{el}}^{\text{OK}}$) were also calculated to evaluate the accuracy of the DFT calculation of the energy involved in proton transfers between side chains, an extremely common situation in, e.g., reactions catalyzed by enzymes. $\Delta \text{PA}_{\text{el}}^{\text{OK}}$ may eventually benefit from the cancellation of the systematic errors resulting from the calculation of $\text{PA}_{\text{el}}^{\text{OK}}$ and is the quantity of interest in many mechanistic studies.

DFT Calculations. As the accuracy of DFT strongly depends upon the choice of the functional used for the specific system and property, single point energy calculations with 64 density functionals (2 of the local spin density approximation (LSDA) type, 16 of the generalized gradient approximation (GGA) type, 8 of the meta-GGA (m-GGA) type, 21 of the hybrid-GGA (h-GGA) type, 14 of the hybrid-meta GGA (hm-GGA) type, and 3 double hybrid-GGA (hh-GGA) type; please see Table 5 for a complete description of the functionals used) were carried out using the 6-311++G(2d,2p) basis set. We have chosen the 6-311++G(2d,2p) basis set because it was the most complete basis set that can be used with most biological systems that typically include over 100 atoms. Calculations with more complete basis sets would become intolerably time-consuming in large systems. Moreover, this basis set is usually close to the CBS limit with DFT, which converges rather quickly with basis set size,⁴⁰ and differences in energy arising from the use of larger basis sets usually amount to a few tenths of kilocalorie per mole. The accuracy of DFT results is thus determined by both the truncation error of the basis set employed and the intrinsic error due to the approximate nature of each functional.

Post-HF and DFT calculations were carried out using the Gaussian 03 suite of programs,⁴¹ except for the double hybrids and dispersion-corrected (B3LYP-D and B97-D) functionals that were carried out using the Gaussian 09 suite of programs.⁴² We have used the default grid sizes. The meta-hybrid density functionals are mostly sensitive to the grid size (in particular the M06 suite, which is known to be extremely sensitive in this regard).⁴³ To check whether the default grid was accurate enough, we tested the default grid (a pruned (75,302) grid) and the ultrafine grid (a pruned (99 590) grid). We have repeated the calculations with the M06 and M06-2X functionals, and the largest difference obtained was only 0.03 kcal/mol in $\text{PA}_{\text{el}}^{\text{OK}}$ (average absolute difference of 0.02 kcal/mol). The influence of the grid size used was found to be insignificant on the systems and properties studied here.

Grimme's D3 correction⁴⁴ was applied to a few density functionals (B3LYP-D3, MPW1B95-D3, B1B95-D3, CAM-B3LYP-D3, and BMK-D3). The D3 correction was calculated using the program available online at Grimme's Web site⁴⁵ and using the Becke–Johnson damping.^{46–48}

RESULTS AND DISCUSSION

Calculation of $\text{PA}_{\text{el}}^{\text{OK}}$ at the MP2/CBS Level. There are few methods to accurately determine thermochemical parameters, and the extrapolation to the complete basis set (CBS) using a

Table 1. PA_{el}^{OK} at the MP2 Level and the Correlation-Consistent aug-cc-pVDZ, aug-cc-pVTZ, and aug-cc-pVQZ Basis Sets (All Values in kcal/mol)

amino acid	PA_{el}^{OK} for each basis set		
	aug-cc-pVDZ	aug-cc-pVTZ	aug-cc-pVQZ
Asp ⁻	346.36	347.50	347.42
Cys ⁻	351.94	353.92	353.87
His	235.29	236.23	236.27
Ser ⁻	374.11	375.49	375.53
Lys	226.33	226.85	226.76
Arg	245.96	247.50	247.71
Tyr ⁻	352.49	354.21	354.36

series of correlation consistent bases together with high-level post-HF methods is among those that provide the best results. Here, we have used high-level post-HF *ab initio* methods (MP2 and CCSD(T)), which were corrected using basis set extrapolation techniques to obtain reference CCSD(T)/CBS energies. Two extrapolation schemes were used and compared. Subsequently, we performed a comparison between the results of 64 DFT functionals and the extrapolated post-HF *ab initio* methods to examine the performance of various DFT functionals in predicting the PA_{el}^{OK} values for every ionizable amino acid side chain. The absolute electronic energies calculated at the HF/aug-cc-pVXZ ($X = 2-4$), MP2/aug-cc-pVXZ ($X = 2-4$), and CCSD(T)/aug-cc-pVXZ ($X = 2-3$) theoretical levels can be found in Table S1 of Supporting Information. Table 1 shows the PA_{el}^{OK} values obtained at the MP2 level and the aug-cc-pVDZ, aug-cc-pVTZ, and aug-cc-pVQZ basis sets.

Table 2 summarizes the PA_{el}^{OK} values at the MP2/CBS level (the separate HF and correlation energies can be found in Table S2 of the Supporting Information) obtained with the extrapolation scheme of Truhlar. The differences between PA_{el}^{OK} at the MP2/CBS and MP2/aug-cc-pVXZ ($X = 2-4$) levels are also shown. The results reveal that the MP2/CBS values are not closer to the quadruple- ζ result than to the triple- ζ result (both the MUE and the MSE to the CBS values are equivalent for the triple and quadruple basis sets). The convergence of PA_{el}^{OK} with the basis set size is impressively fast. The results with the aug-cc-pVTZ basis set are already at 0.17 kcal/mol from the CBS limit (in average), and the maximum difference between the triple- ζ and the CBS values is 0.29 kcal/mol. We conclude that the extrapolated values for these systems and property are not obviously more accurate than the triple- ζ values. However, this conclusion only holds for PA_{el}^{OK} , which is based on energy difference. If we consider the absolute electronic energies of the protonated and deprotonated species, then the extrapolated values are much closer to the quadruple- ζ results (average MUE = 0.08 kcal/mol, average MSE = -0.08 kcal/mol) than to the triple- ζ results (average MUE = 0.18 kcal/mol, average MSE = -0.18 kcal/mol). Further details can be found in Tables S2 and S3 in the Supporting Information.

We have calculated also the PA_{el}^{OK} at the MP2/CBS level using the Helgaker extrapolation scheme. This extrapolation involved the MP2/aug-cc-pVTZ and MP2/aug-cc-pVQZ electronic energies, as well as the Truhlar extrapolated HF energies. The values provided by this method are expected to be more accurate and reliable because they are based in larger basis sets than those used in the Truhlar method (they are also

Table 2. PA_{el}^{OK} Calculated at the MP2/CBS Level Using the Truhlar Extrapolation Scheme^a

amino acids	PA_{el}^{OK}			
	MP2/CBS	MP2/CBS-MP2/aug-cc-pVDZ	MP2/CBS-MP2/aug-cc-pVTZ	MP2/CBS-MP2/aug-cc-pVQZ
Asp ⁻	347.56	1.20	0.06	0.14
Cys ⁻	354.08	2.14	0.15	0.20
His	235.99	0.70	-0.24	-0.28
Ser ⁻	375.68	1.57	0.19	0.15
Lys	226.71	0.38	-0.15	-0.05
Arg	247.42	1.45	-0.09	-0.30
Tyr ⁻	354.51	2.02	0.29	0.14
MUE		1.35	0.17	0.18
MSE		1.35	0.03	0.00

^a The difference between PA_{el}^{OK} values obtained with a CBS and with the aug-cc-pVXZ ($X = 2-4$) is also shown. All values in kcal/mol.

Table 3. PA_{el}^{OK} Values at the MP2/CBS Level Obtained with the Helgaker Extrapolation Scheme Using the aug-cc-pVTZ and aug-cc-pVQZ Basis Sets, As Well As the Difference between the Truhlar and Helgaker et al. Extrapolation Methods (ΔPA_{el}^{OK} ; All Values in kcal/mol)

amino acids	PA_{el}^{OK} (MP2/CBS)	ΔPA_{el}^{OK}
Asp ⁻	347.19	0.38
Cys ⁻	353.22	0.86
His	236.27	-0.28
Ser ⁻	375.37	0.31
Lys	226.69	0.02
Arg	247.84	-0.43
Tyr ⁻	354.31	0.20

much more expensive in terms of CPU power and seldom feasible in biological systems of interest with up to 100 atoms, in which the use of the aug-cc-pVTZ basis set is already a challenge). Table 3 shows the MP2/CBS PA_{el}^{OK} values obtained with the Helgaker method, as well as the difference between both extrapolated methodologies.

The data shows that the difference between the PA_{el}^{OK} values at the MP2/CBS level extrapolated with both methods is very small in most cases (absolute average of 0.35 kcal/mol only). The only exception is Cys, which shows a more significant deviation (-0.86 kcal/mol). Even though the use of such large basis sets might be very useful in applications involving a very limited number of atoms/electrons, if we consider typical (larger) biological systems, the extrapolation using quadruple- ζ basis sets will probably be detrimental for the quality of the overall results. The application of this large basis set would force the researcher to reduce significantly the size of the molecular models, introducing errors much larger than the few tenths of a kilocalorie per mole that can result in terms of accuracy (assuming that the difference between both methods is translated into a gain in accuracy for the extrapolation using larger basis sets).

Calculation of PA_{el}^{OK} at the CCSD(T)/CBS Level. The CCSD(T)/CBS energies were obtained adding to the MP2/CBS energies the energy difference between CCSD(T) and MP2

Table 4. $\Delta E_{\text{CCSD(T)}-\text{MP2}}$ with the Pople Basis Sets (6-31G*(0.25), 6-31+G(d), and 6-31+G(d,p)) and Correlation-Consistent Basis Sets (aug-cc-pVDZ and aug-cc-pVTZ)^a

amino acids	$\Delta E_{\text{CCSD(T)}-\text{MP2}}$					$\text{PA}_{\text{el}}^{\text{OK}}$
	6-31G*(0.25)	6-31+G(d)	6-31+G(d,p)	aug-cc-pVDZ	aug-cc-pVTZ	CCSD(T)/CBS
Asp ⁻	1.23	2.58	2.76	2.60	2.61	349.80
Cys ⁻	2.29	1.72	1.77	2.81	2.34	355.55
His	2.39	2.35	2.40	2.43	2.58	238.84
Ser ⁻	3.20	2.30	2.53	2.37	2.53	377.90
Lys	1.27	1.09	1.26	1.18	1.34	228.03
Arg	2.68	1.18	1.39	1.40	1.62	249.47
Tyr ⁻	3.28	3.17	3.33	3.05	3.05	357.36
MUE	0.52	0.27	0.21	0.17	0.00	
MaxE	1.38	0.62	0.57	0.47	0.00	

^a The mean unsigned error (MUE) and maximum error (MaxE) between the $\Delta E_{\text{CCSD(T)}-\text{MP2}}$ values of each basis set and the largest basis set (aug-cc-pVTZ) are also shown. The $\Delta E_{\text{CCSD(T)}-\text{MP2}}$ value with the aug-cc-pVTZ was used to calculate the $\text{PA}_{\text{el}}^{\text{OK}}$ for each amino acid extrapolated to the CCSD(T)/CBS level (last column on the right of the table). All values in kcal/mol.

($\Delta E_{\text{CCSD(T)}-\text{MP2}}$). This last term was calculated with several basis sets to check for its convergence with basis set size. The $\Delta E_{\text{CCSD}-\text{MP2}}$ and $\text{PA}_{\text{el}}^{\text{OK}}$ values obtained in this way are shown in Tables 4 and 5.

The difference between CCSD(T) and MP2 is significant, as expected. However, such a difference seems to be efficiently accounted for by the $\Delta E_{\text{CCSD}-\text{MP2}}$ term with the aug-cc-pVTZ basis set. The average MUE $\Delta E_{\text{CCSD}-\text{MP2}}$ difference between the double and triple- ζ basis sets is only 0.17 kcal/mol, and the MaxE is 0.47 kcal/mol. We cannot calculate exactly the accuracy of the $\Delta E_{\text{CCSD}-\text{MP2}}$ term because we would need to know the CCSD(T)/CBS energy extrapolated from the CCSD(T) energies only. As we have only the double- and triple- ζ basis sets results with this method, we cannot obtain an accurate extrapolation of the CCSD(T) just from them. However, the data in Table 4 clearly show that the accuracy of $\Delta E_{\text{CCSD}-\text{MP2}}$ increases monotonically with basis set size (both MUE and MaxE) and that the 6-31G*(0.25) basis set should be avoided in these systems as the deviation for the largest basis set is significantly large. The aug-cc-pVTZ basis set should be used, at least to obtain predictions that safely converge within chemical accuracy. However, this basis set is prohibitively large for many applications at the CCSD(T) level.

On the basis of these observations, we have used the $\Delta E_{\text{CCSD}-\text{MP2}}$ correction with the aug-cc-pVTZ basis set and the MP2/CBS values obtained with the Helgaker et al. extrapolation scheme to obtain the $\text{PA}_{\text{el}}^{\text{OK}}$ at the CCSD(T)/CBS level.

The MP2/CBS extrapolated energy differences converge much better and faster than absolute energies but are still approximations. The frozen-core approximation and the conversion of MP2/CBS into CCSD(T)/CBS through eq 7 also introduce small errors in the final values. Others sources of error (e.g., the use of aug-cc-pVTZ to describe the sulfur atom or the truncation of the expansion of the cluster operator) are present as well. Altogether, these values may result in an uncertainty of a few tenths of a kilocalorie per mole (the exact value is difficult to estimate). As these approximated CCSD(T)/CBS energy differences were used to benchmark the density functionals, a part of the error attributed to DFT might arise from the uncertainty in the reference CCSD(T)/CBS values, which may slightly affect the order in which the functionals are classified.

Benchmarking of Density Functionals. This assessment does not represent the global quality of the functionals, which must be measured through the calculation of diverse sets of properties in a representative set of molecular systems. Instead, this study just evaluates specifically the accuracy of the description of the zero-point exclusive proton affinity at 0 K for amino acid side chains. We have benchmarked 64 density functionals (2 LSDA, 16 GGA, 8 m-GGA, 21 h-GGA, 14 hm-GGA, and 3 hh-GGA; please see Table 5 for a complete description of the functionals used) through single point energy calculations at the DFT/6-311++G(2d,2p)//MP 2/6-311++G(d,p) level. The mean unsigned error (MUE $\text{PA}_{\text{el}}^{\text{OK}}$) and maximum error (MaxE $\text{PA}_{\text{el}}^{\text{OK}}$) between these values and the CCSD(T)/CBS value is shown in Table 5.

All of these values are important in order to understand the accuracy provided by each functional in, e.g., biological acid/base catalysis, in which some of these amino acids are involved in the catalytic cycles. We have divided the density functionals into three groups according to the $\Delta \text{PA}_{\text{el}}^{\text{OK}}$ MUE. In group I, the MUE ranges from 0 to 1.42 kcal/mol (which corresponds to an error below 1 pK_a unit at 310.15 K). Group II includes the functionals with MUE between 1.42 and 2.84 kcal/mol (which corresponds to an error between 1 and 2 pK_a units at 310.15 K). Group III includes functionals with a MUE above 2.84 kcal/mol (errors above 2 pK_a units). The density functionals that provide the more accurate $\text{PA}_{\text{el}}^{\text{OK}}$ values are MPW1B95-D3, XYG3, MPW1B95, B1B95-D3, BMK, BMK-D3, M06-2X, and B1LYP. They have a MUE close to chemical accuracy and MaxE of circa 2 kcal/mol. The remaining functionals of group I (B1B95, PBE1PBE, CAM-B3LYP, B97-1, PBE1KCIS, B3P86, CAM-B3LYP-D3, B3LYP, B98, and M06-L) and both M06 and B3LYP-D (marginally in group II but better than any other of this group) also show very satisfactory performance with MUEs below 1.48 kcal/mol and MaxE's of circa 3 kcal/mol. MPW1B95 in particular has the lowest MUE and one of the lowest MaxE's. All of these functionals are nonlocal, even though there is no correlation between the MUE and the fraction of HF exchange. The very popular B3LYP functional ranks 16th in terms of MUE and 11th in terms of MaxE, belonging to the group of the most accurate functionals. The performance of the dispersion-corrected B3LYP-D functional ranks 19th and 18th in terms of MUE and MaxE.

Table 5. Differences in the PA_{el}^{OK} Values Calculated at the DFT/6-311++G(2d,2p)//MP2/6-311++G(d,p) Level and at the CCSD(T)/CBS Level, for Each Amino Acid^a

functional	type	%HF χ	MUE	MaxE	Asp ⁻	Cys ⁻	His	Ser ⁻	Arg	Lys	Tyr ⁻
MPW1B95-D3 ^{44,49-52}	hm-GGA	31	0.79	2.25	-0.24	0.50	1.05	-0.72	2.25	0.55	-0.24
XYG3 ^{53,54}	hh-GGA	80.33	0.83	1.87	-1.04	0.70	-0.50	-1.28	0.15	0.29	-1.87
MPW1B95 ⁴⁹⁻⁵²	hm-GGA	31	0.85	2.49	-0.03	0.78	1.29	-0.56	2.49	0.78	-0.03
B1B95-D3 ^{44,50,55}	hm-GGA	28	0.91	2.29	-0.13	0.67	1.66	-0.73	2.29	0.66	-0.25
BMK ⁵⁶	hm-GGA	42	0.93	1.83	-0.72	-1.66	1.06	0.10	1.83	0.25	-0.90
BMK-D3 ^{44,56}	hm-GGA	42	1.00	2.21	-1.10	-2.21	0.60	-0.19	1.40	-0.20	-1.29
M06-2X ^{17,57}	hm-GGA	54	1.08	2.09	-0.47	-2.09	-1.64	-0.25	-0.81	-1.52	-0.78
B1LYP ^{55,58,59}	h-GGA	25	1.10	1.92	-1.17	-0.06	1.27	-1.92	1.88	0.34	-1.05
B1B95 ^{50,55}	hm-GGA	28	1.11	2.77	0.28	1.28	1.66	-0.40	2.77	1.16	0.18
PBE1PBE ⁶⁰	h-GGA	25	1.17	2.98	0.17	0.79	1.79	-0.69	2.98	1.39	-0.40
CAM-B3LYP ⁶¹	h-GGA	19/65**	1.18	2.00	-2.00	-0.57	0.25	-1.72	1.18	-0.99	-1.51
B97-1 ⁶²	h-GGA	21	1.31	2.91	0.24	0.73	2.23	-0.96	2.91	1.81	-0.26
PBE1KCIS ^{60,63,64}	hm-GGA	22	1.31	2.85	-0.24	0.93	1.96	-1.31	2.85	1.30	-0.57
B3P86 ^{50,51,55}	h-GGA	20	1.32	3.11	0.03	1.21	2.13	-0.95	3.11	1.49	-0.32
CAM-B3LYP-D3 ^{44,61}	h-GGA	19/65**	1.33	2.30	-2.30	-0.96	-0.11	-1.96	0.83	-1.35	-1.82
B3LYP ^{50,55,59}	h-GGA	20	1.34	2.78	-1.60	-0.18	1.21	-2.78	1.78	0.22	-1.61
B98 ⁶⁵	h-GGA	21.98	1.35	3.17	0.44	0.79	2.47	-0.57	3.17	1.95	0.06
M06-L ⁶⁶	m-GGA	0	1.38	2.95	1.14	2.95	1.27	-1.19	2.26	0.56	-0.30
B3LYP-D ⁶⁷⁻⁶⁹	h-GGA	20	1.48	3.00	-1.08	0.29	1.87	-2.13	3.00	1.04	-0.93
M06 ^{17,57}	hm-GGA	27	1.48	2.95	-1.42	-0.06	-1.09	-2.95	0.32	-2.26	-2.30
B3LYP-D3 ^{44,50,55,59}	h-GGA	20	1.56	3.28	-2.23	-1.05	0.42	-3.28	1.01	-0.60	-2.29
B3PW91 ^{50,55,70}	h-GGA	20	1.63	3.68	0.58	1.89	2.67	-0.42	3.68	2.08	0.11
MPW3LYP ^{49,59,70}	h-GGA	21.8	1.65	3.67	-2.35	-0.96	0.63	-3.67	1.19	-0.39	-2.35
MPW1KCIS ^{16,49,64,70}	hm-GGA	15	1.66	2.85	-0.60	1.22	2.19	-2.26	2.85	1.49	-1.02
MPW1PW91 ^{49,70}	h-GGA	25	1.72	3.77	0.88	1.90	2.61	0.25	3.77	2.14	0.51
TPSS1KCIS ^{64,71,72}	hm-GGA	13	1.75	2.84	-0.48	1.46	2.32	-2.26	2.84	2.14	-0.75
O3LYP ^{59,73,74}	h-GGA	11.61	1.86	4.01	0.77	2.35	3.09	0.06	4.01	2.73	-0.03
TPSSH ⁷¹	hm-GGA	10	1.89	3.12	-0.14	1.96	2.79	-2.02	3.12	2.81	-0.40
MPW1S ^{49,70,75}	h-GGA	^b	1.97	4.91	-2.11	-0.05	1.28	-4.91	1.82	0.69	-2.95
VSXC ⁷⁶	m-GGA	0	2.10	5.16	-0.97	0.51	3.27	-5.16	3.05	1.47	-0.27
BHandH ^{55,59}	h-GGA	50	2.15	4.19	-2.52	-4.19	-1.31	-1.81	0.66	-2.26	-2.29
BPBE ^{55,60}	GGA	0	2.23	5.55	-2.13	0.23	1.56	-5.55	1.93	1.04	-3.18
G96LYP ^{59,77}	GGA	0	2.25	6.07	-2.88	-0.36	1.40	-6.07	1.35	0.36	-3.33
MPWB1K ⁷⁸	hm-GGA	44	2.26	3.78	1.73	1.85	2.19	2.31	3.78	1.76	2.16
BPW91 ^{55,70}	GGA	0	2.26	5.46	-2.08	0.39	1.68	-5.46	2.04	1.10	-3.04
HCTH147 ⁶²	GGA	0	2.27	3.92	-0.53	1.39	2.87	-3.92	3.44	2.26	-1.46
TPSSTPSS ⁷¹	m-GGA	0	2.33	4.71	-1.65	1.09	2.28	-4.71	2.24	2.32	-2.03
OLYP ^{59,73}	GGA	0	2.34	4.39	-0.58	1.42	2.63	-4.39	3.23	2.42	-1.74
HCTH407 ⁶²	GGA	0	2.41	4.10	-0.38	1.31	3.11	-4.08	4.10	2.49	-1.43
TPSSLYP1W ^{59,71,79}	m-GGA	0	2.42	6.08	-3.00	0.29	2.00	-6.08	1.30	1.61	-2.62
B97-D ⁶⁸	GGA	0	2.43	4.30	-0.30	1.55	3.44	-3.46	4.30	3.04	-0.89
MPWKCIS ^{49,64,70}	m-GGA	0	2.45	6.81	-3.20	-0.58	1.10	-6.81	1.25	0.31	-3.90
MPW91 ^{49,70}	GGA	0	2.45	6.81	-3.19	-0.80	0.85	-6.81	1.18	0.22	-4.11
BB1K ^{50,55,80}	hm-GGA	42	2.56	4.04	2.06	2.29	2.50	2.57	4.04	2.08	2.40
PBE1W ^{60,79}	GGA	0	2.66	7.52	-3.73	-1.46	0.42	-7.52	0.63	-0.27	-4.60
wB97X-D ^{81,82}	h-GGA	22.20/100 ^c	2.67	4.82	1.59	2.39	3.55	1.85	4.82	2.83	1.67
BP86 ^{51,55}	GGA	0	2.70	7.48	-3.83	-1.96	0.01	-7.48	0.15	-0.55	-4.88
HCTH93 ⁶²	GGA	0	2.70	4.79	0.99	3.09	4.08	-2.18	4.79	3.62	-0.11
BB95 ^{50,55}	m-GGA	0	2.74	8.01	-4.03	-1.43	-0.07	-8.01	0.12	-0.75	-4.75
BHandHLYP ^{55,59}	h-GGA	50	2.74	4.26	1.98	1.82	2.88	3.29	4.26	2.05	2.91
B2PLYP ⁸³	hh-GGA	53	2.78	4.13	2.18	1.63	2.87	3.51	4.13	2.19	2.94
B97-2 ⁶²	h-GGA	21	2.86	4.78	2.10	3.28	3.71	1.21	4.78	3.34	1.65
PBEPBE ⁶⁰	GGA	0	3.02	8.15	-4.22	-2.28	-0.23	-8.15	0.12	-0.79	-5.34

Table 5. Continued

functional	type	%HF χ	MUE	MaxE	Asp ⁻	Cys ⁻	His	Ser ⁻	Arg	Lys	Tyr ⁻
mPW2PLYP ⁸⁴	hh-GGA	55	3.28	5.17	2.48	1.79	3.30	4.00	5.17	2.75	3.47
PBELYP1W ^{59,60,79}	GGA	0	3.29	8.83	-4.94	-2.37	-0.11	-8.83	-0.34	-1.19	-5.30
BLYP ^{55,59}	GGA	0	3.52	8.87	-5.10	-2.69	-0.38	-8.87	-0.63	-1.44	-5.55
MPW1N ^{49,70,85}	h-GGA	<i>b</i>	3.62	5.33	3.00	3.20	3.68	3.73	5.33	3.29	3.11
MPWKIS1K ^{16,49,64,70}	hm-GGA	41	3.86	5.49	3.11	3.57	4.01	3.96	5.49	3.44	3.47
MPW1K ^{49,70,75,86}	h-GGA	42.8	3.87	5.54	3.28	3.37	3.83	4.18	5.54	3.45	3.46
MPWLYP1W ^{49,59,70,79}	GGA	0	3.95	9.52	-5.60	-3.16	-0.68	-9.52	-0.93	-1.77	-5.99
MPWB95 ^{49,50,70}	m-GGA	0	4.05	9.37	-5.14	-2.65	-0.92	-9.37	-2.81	-1.63	-5.84
MPWLYP ^{49,59,70}	GGA	0	4.57	10.23	-6.19	-3.92	-1.22	-10.23	-1.48	-2.31	-6.62
SVWN3 ^{87,88}	LSDA	0	7.45	13.08	-9.02	-8.00	-3.84	-13.08	-2.79	-5.36	-10.03
SVWN5 ^{87,88}	LSDA	0	8.25	14.04	-9.74	-9.07	-4.49	-14.04	-3.54	-5.97	-10.91

^a The mean unsigned error (MUE PA_{el}^{OK}) and maximum error (MaxE PA_{el}^{OK}) are also shown. All values in kcal/mol. ^b Not available. ^c The first value is %HF χ at short range, and the second value is %HF χ at long range.

The inclusion of dispersion effects (D correction) in B3LYP systematically increased the PA value and worsened the overall results (+0.14 kcal/mol in the MUE and +0.23 kcal/mol in the MaxE). The opposite tendency was seen with the D3 correction. Here, the correction, in all of the five tested functionals and in every amino acid, decreased the PA values by 0.2–0.7 kcal/mol. In three functionals, the D3 correction also worsened the results (B3LYP-D3, +0.22 in MUE and +0.50 in MaxE; CAM-B3LYP-D3, +0.16 in MUE and +0.30 in MaxE; BMK-D3, +0.07 in MUE and +0.37 in MaxE), while in the other two functionals, the results were improved (MPW1B95-D3, -0.06 in MUE and -0.24 in MaxE; B1B95-D3, -0.19 in MUE and -0.48 in MaxE). In general, the influence of the dispersion corrections in these systems is small (always below 1.2 kcal/mol, 0.5 kcal/mol in average). As we deal with single, small molecules, there are few dispersive interactions in the zero overlap region, where dispersion is unaccounted for by density functionals. Moreover, as the geometries of the protonated and unprotonated amino acids are very similar, and the difference is just a proton, most of the dispersion cancels out when we calculate PA_{el}^{OK}.

In general, there is a positive correlation between MUE and MaxE. The MaxE for each density functional is not always provided by the same amino acid. If we consider only groups I and II (MUE < 2.84 kcal/mol, pK_a errors < 2), we can see that there are no significant differences in the MUEs of each amino acid (averaged over all of those functionals), except for serine, in which the MUE is larger on average (MUE(Asp⁻) = 1.4 kcal/mol, MUE(Cys⁻) = 1.3 kcal/mol, MUE(His) = 1.8 kcal/mol, MUE(Ser⁻) = 2.9 kcal/mol, MUE(Arg) = 2.4 kcal/mol, MUE(Lys) = 1.4 kcal/mol, MUE(Tyr⁻) = 1.6 kcal/mol). With very few exceptions M06, M06-2X, CAM-B3LYP-D3, CAM-B3LYP, and BMK-D3 systematically underestimated PA_{el}^{OK} while B98, B1B95, B97-1, B3P86, and M06-L overestimated this value.

In recent papers, it was suggested that M06-2X and M06 were the best functionals for a combination of main-group thermochemistry, kinetics, and noncovalent interactions.^{17,57} Other studies have concluded that B3LYP and B98 were very precise in the calculation of proton affinities for small models such as pyrrole, quinoline,⁸⁹ γ -butyrolactone, and 2-pyrrolidinone.¹⁴ Our results show that all of these hybrid and hybrid-meta functionals are very good to describe the PA_{el}^{OK} of the ionizable amino acid side chains, with errors below 1 pK_a unit.

Many biological mechanisms involve proton transfers between ionizable amino acid side chains (e.g., reaction mechanisms

in enzymes). Therefore, it becomes very important to evaluate the performance of the density functionals in systems that involve proton transfers between different amino acids. For that purpose, we have calculated the accuracy of the proton transfer energies (ΔE_{PT}) between every pair of the seven studied amino acids for the 20 more accurate density functionals (MPW1B95, PBE1KIS, M06-2X, B1LYP, B1B95, PBE1PBE, B97-1, B3P86, B3LYP, B98, M06, M06-L, MPW1B95-D3, B1B95-D3, BMK, BMK-D3, CAM-B3LYP, CAM-B3LYP-D3, B3LYP-D, and XYG3). The accuracy was measured by the MUE between the ΔE_{PT} energies calculated at the DFT level and at the CCSD(T)/CBS level.

Table 6 shows the errors (differences between the DFT and CCSD(T)/CBS energies) obtained for each proton transfer between every amino acid pair. If systematic errors are present in PA_{el}^{OK}, the accuracy of ΔE_{PT} will be greater than the accuracy of PA_{el}^{OK}. In the absence of systematic errors, the error should be larger in ΔE_{PT} than in PA_{el}^{OK} (due to error propagation). Table 7 shows the MUE in ΔE_{PT} for the 20 density functionals.

The present results can be used in many ways. To begin with, they clearly show that M06-2X is the most accurate density functional for proton transfers, in general, and B3LYP and B3LYP-D are the less accurate, in general, among the 20 best rated to predict PA_{el}^{OK} values. On the other hand, if one wants to be very specific, the ΔE_{PT} values are useful in identifying the best density functional to describe each specific enzymatic system. For example, for the pancreatic elastase catalytic mechanism⁹⁰ in which a proton transfer occurs from one serine to one histidine, we see that the three density functionals that provide the most accurate results are XYG3, BMK-D3, BMK, and M06-2X. Therefore, one of these four functionals might be used in proteins belonging to the serine protease family. Similarly, the results suggest that the M06-L, M06, XYG3, and M06-2X density functionals are the best to describe the proton transfer from carboxylic acids to histidines that occurs in the sulfotransferase enzymes.⁹¹ However, we must keep in mind that the energetics of enzyme reactions depends on many more factors beyond proton transfers and that the choice of the density functional should not be only grounded in this criterion.

It is important to note that the pK_a values of the ionizable residues in folded proteins are influenced by the local environment. Therefore, the results presented in this work show only how well DFT describes the gas-phase proton affinities. The proton affinities in proteins will also depend on the description of

noncovalent interactions (among other factors), which are different for each functional.

Table 6. ΔE_{PT} for the 12 Density Functionals^a

		MPW1B95						
		Asp ⁻	Cys ⁻	His	Ser ⁻	Arg	Lys	Tyr ⁻
PBE1KCIS	Asp ⁻	0	0.81	1.33	-0.52	2.52	0.82	0.01
	Cys ⁻	1.16	0	0.52	-1.34	1.71	0.01	-0.81
	His	2.19	1.03	0	-1.85	1.20	-0.51	-1.32
	Ser ⁻	-1.08	-2.24	-3.27	0	3.05	1.34	0.53
	Arg	3.09	1.92	0.89	4.16	0	-1.70	-2.52
	Lys	1.54	0.38	-0.65	2.62	-1.55	0	-0.81
	Tyr ⁻	-0.33	-1.50	-2.53	0.74	-3.42	-1.88	0
		M062X						
		Asp ⁻	Cys ⁻	His	Ser ⁻	Arg	Lys	Tyr ⁻
B1LYP	Asp ⁻	0	-1.62	-1.17	0.22	-0.34	-1.05	-0.31
	Cys ⁻	1.11	0	0.45	1.84	1.28	0.57	1.31
	His	2.44	1.33	0	1.39	0.83	0.12	0.86
	Ser ⁻	-0.75	-1.86	-3.19	0	-0.56	-1.27	-0.53
	Arg	3.05	1.94	0.61	3.80	0	-0.71	0.03
	Lys	1.51	0.40	-0.93	2.26	-1.53	0	0.74
	Tyr ⁻	0.12	-0.99	-2.32	0.87	-2.93	-1.39	0
		B1B95						
		Asp ⁻	Cys ⁻	His	Ser ⁻	Arg	Lys	Tyr ⁻
PBE1PBE	Asp ⁻	0	1.00	1.37	-0.69	2.49	0.88	-0.10
	Cys ⁻	0.62	0	0.38	-1.69	1.49	-0.12	-1.10
	His	1.62	0.99	0	-2.06	1.11	-0.50	-1.47
	Ser ⁻	-0.86	-1.48	-2.48	0	3.18	1.57	0.59
	Arg	2.81	2.19	1.19	3.67	0	-1.61	-2.59
	Lys	1.22	0.59	-0.40	2.08	-1.59	0	-0.98
	Tyr ⁻	-0.58	-1.20	-2.19	0.28	-3.39	-1.79	0
		B97-1						
		Asp ⁻	Cys ⁻	His	Ser ⁻	Arg	Lys	Tyr ⁻
B3P86	Asp ⁻	0	0.50	1.99	-1.20	2.67	1.57	-0.50
	Cys ⁻	1.18	0	1.50	-1.69	2.17	1.08	-0.99
	His	2.09	0.92	0	-3.19	0.68	-0.42	-2.49
	Ser ⁻	-0.99	-2.16	-3.08	0	3.87	2.77	0.70
	Arg	3.08	1.90	0.99	4.07	0	-1.10	-3.17
	Lys	1.46	0.28	-0.64	2.44	-1.62	0	-2.07
	Tyr ⁻	-0.35	-1.53	-2.44	0.64	-3.43	-1.81	0
		B3LYP						
		Asp ⁻	Cys ⁻	His	Ser ⁻	Arg	Lys	Tyr ⁻
B98	Asp ⁻	0	1.43	2.81	-1.17	3.38	1.82	-0.01
	Cys ⁻	0.35	0	1.39	-2.60	1.96	0.40	-1.44
	His	2.03	1.68	0	-3.99	0.57	-0.99	-2.82
	Ser ⁻	-1.01	-1.36	-3.04	0	4.56	3.00	1.17
	Arg	2.73	2.38	0.70	3.74	0	-1.56	-3.39
	Lys	1.51	1.16	-0.52	2.52	-1.22	0	-1.83
	Tyr ⁻	-0.38	-0.73	-2.41	0.63	-3.11	-1.88	0

Table 6. Continued

		M06						
		Asp ⁻	Cys ⁻	His	Ser ⁻	Arg	Lys	Tyr ⁻
M06-L	Asp ⁻	0	1.36	0.33	-1.53	1.74	-0.84	-0.88
	Cys ⁻	1.81	0	-1.03	-2.89	0.38	-2.20	-2.24
	His	0.13	-1.68	0	-1.86	1.41	-1.17	-1.21
	Ser ⁻	-2.33	-4.13	-2.46	0	3.27	0.69	0.65
	Arg	1.12	-0.69	0.99	3.45	0	-2.58	-2.62
	Lys	-0.58	-2.38	-0.71	1.75	-1.69	0	-0.04
	Tyr ⁻	-1.44	-3.25	-1.57	0.89	-2.56	-0.86	0
		MPW1B95-D3						
		Asp ⁻	Cys ⁻	His	Ser ⁻	Arg	Lys	Tyr ⁻
B1B95-D3	Asp ⁻	0	0.75	1.29	-0.48	2.49	0.79	0.00
	Cys ⁻	0.80	0	0.55	-1.23	1.75	0.04	-0.74
	His	1.79	0.98	0	-1.78	1.20	-0.50	-1.29
	Ser ⁻	-0.60	-1.40	-2.38	0	2.98	1.27	0.49
	Arg	2.42	1.62	0.64	3.02	0	-1.71	-2.49
	Lys	0.79	-0.01	-1.00	1.38	-1.63	0	-0.79
	Tyr ⁻	-0.12	-0.92	-1.91	0.47	-2.54	-0.91	0
		BMK						
		Asp ⁻	Cys ⁻	His	Ser ⁻	Arg	Lys	Tyr ⁻
BMK-D3	Asp ⁻	0	-0.94	1.78	0.82	2.55	0.97	-0.18
	Cys ⁻	-1.11	0	2.71	1.76	3.49	1.90	0.76
	His	1.70	2.81	0	-0.96	0.78	-0.81	-1.95
	Ser ⁻	0.91	2.01	-0.80	0	1.73	0.15	-1.00
	Arg	2.50	3.61	0.80	1.60	0	-1.59	-2.73
	Lys	0.90	2.01	-0.80	-0.01	-1.60	0	-1.14
	Tyr ⁻	-0.19	0.92	-1.89	-1.09	-2.69	-1.09	0
		CAM-B3LYP						
		Asp ⁻	Cys ⁻	His	Ser ⁻	Arg	Lys	Tyr ⁻
CAM-B3LYP-D3	Asp ⁻	0	1.43	2.25	0.28	3.18	1.01	0.49
	Cys ⁻	1.35	0	0.83	-1.14	1.76	-0.42	-0.93
	His	2.20	0.85	0	-1.97	0.93	-1.25	-1.76
	Ser ⁻	0.34	-1.00	-1.85	0	2.90	0.73	0.21
	Arg	3.13	1.79	0.94	2.79	0	-2.18	-2.69
	Lys	0.95	-0.39	-1.24	0.61	-2.18	0	-0.51
	Tyr ⁻	0.49	-0.86	-1.71	0.14	-2.65	-0.47	0
		XYG3						
		Asp ⁻	Cys ⁻	His	Ser ⁻	Arg	Lys	Tyr ⁻
B3LYP-D	Asp ⁻	0	1.74	0.55	-0.24	1.19	1.33	-0.83
	Cys ⁻	1.37	0	-1.19	-1.98	-0.55	-0.41	-2.57
	His	2.95	1.58	0	-0.79	0.64	0.79	-1.37
	Ser ⁻	-1.05	-2.43	-4.01	0	1.43	1.57	-0.59
	Arg	4.08	2.71	1.13	5.14	0	0.14	-2.02
	Lys	2.12	0.75	-0.83	3.17	-1.96	0	-2.16
	Tyr ⁻	0.15	-1.22	-2.80	1.21	-3.93	-1.97	0

^a All possible proton transfers between the seven side chains were included. All values in kcal/mol.

Table 7. MUE of the ΔE_{PT} Averaged over All Proton Transfers with a Given Density Functional (All Values in kcal/mol)

functional	MUE (ΔE_{PT})
M06-2X	0.82
MPW1B95-D3	1.13
XYG3	1.15
MPW1B95	1.20
B1B95	1.28
B1B95-D3	1.30
CAM-B3LYP-D3	1.33
CAM-B3LYP	1.37
BMK	1.46
M06	1.47
BMK-D3	1.48
PBE1PBE	1.58
B98	1.67
B1LYP	1.68
B97-1	1.73
M06-L	1.74
B3P86	1.77
PBE1KCIS	1.82
B3LYP	2.01
B3LYP-D	2.22

CONCLUSIONS

The results presented in this work are important for a conscientious characterization of the protonation states of amino acids and of proton transfers between different amino acids and will be helpful to further studies, e.g., on related acid/base catalytic mechanisms. A new, important and consistent DFT benchmarking database for PA_{el}^{OK} and for ΔPA_{el}^{OK} is provided. The PA_{el}^{OK} reference values were determined at the very accurate CCSD(T)/CBS level. These values were obtained by the sum of the complete basis set limit of the MP2 energies with a CCSD(T) correction term evaluated using the aug-cc-pVTZ basis set. The MP2/CBS energies were determined using both the Truhlar and Helgaker extrapolation schemes. The difference between the PA_{el}^{OK} values at the MP2/CBS level extrapolated with both schemes is very small (absolute average of 0.35 kcal/mol only). Analyzing the results of the DFT benchmarking, we conclude that the M06-2X, XYG3, MPW1B95, and B1B95 functionals are the more adequate to predict amino acid pK_a 's and proton transfers between them. M06-2X is particularly attractive in this regard, as it has been shown to be accurate for intermolecular interactions (which will affect the pK_a in folded proteins) and thermochemistry and kinetics (properties that are fundamental for most biological reactivity studies). It seems clearly to be one of the most attractive choices for studies of chemical reactivity in enzymes in which proton transfers play an important role.

ASSOCIATED CONTENT

S Supporting Information. Absolute electronic energies calculated at the HF/aug-cc-pVXZ ($X = 2-4$), MP2/aug-cc-pVXZ ($X = 2-4$), and CCSD(T)/aug-cc-pVXZ ($X = 2-3$) theoretical levels are shown in Table S1. PA_{el}^{OK} calculated at the

MP2/CBS level using the Truhlar extrapolation scheme as well as the separated HF and correlation energies are shown in Table S2. Mean Unsigned Errors (MUE) and Mean Signed Errors (MSE) for the absolute electronic energy values (total, HF, and Corr) of the difference between the MP2/CBS level using the Truhlar extrapolation scheme and the MP2/aug-cc-pVXZ ($X = 2-4$) basis sets, presented in Tables S3 and S4. Optimized geometries of all species are also presented. This information is available free of charge via the Internet at <http://pubs.acs.org/>.

AUTHOR INFORMATION

Corresponding Author

*E-mail: mjramos@fc.up.pt.

Author Contributions

^SThese authors contributed equally to this work

ACKNOWLEDGMENT

N.F.B. and M.A.S.P. would like to thank the Fundação para a Ciência e a Tecnologia (FCT) for their Post-Doc and Ph.D. grants (SFRH/BPD/71000/2010 and SFRH/BD/43600/2008, respectively). The authors would like to acknowledge the Fundação para a Ciência e a Tecnologia for financial support (PTDC/QUI-QUI/102760/2008).

REFERENCES

- Jordan, I. K.; Kondrashov, F. A.; Adzhubei, I. A.; Wolf, Y. I.; Koonin, E. V.; Kondrashov, A. S.; Sunyaev, S. *Nature* **2005**, *433*, 633.
- Pace, C. N.; Grimsley, G. R.; Scholtz, J. M. *J. Biol. Chem.* **2009**, *284*, 13285.
- Grimsley, G. R.; Scholtz, J. M.; Pace, C. N. *Protein Sci.* **2009**, *18*, 247.
- Harris, T. K.; Turner, G. J. *Iubmb Life* **2002**, *53*, 85.
- Harrison, A. G. *Mass Spectrom. Rev.* **1997**, *16*, 201.
- Matthew, J. B.; Gurd, F. R. N.; Garciamoreno, E. B.; Flanagan, M. A.; March, K. L.; Shire, S. J. *Crc Crit. Rev. Biochem.* **1985**, *18*, 91.
- Moser, A.; Range, K.; York, D. M. *J. Phys. Chem. B* **2010**, *114*, 13911.
- Range, K.; Lopez, C. S.; Moser, A.; York, D. M. *J. Phys. Chem. A* **2006**, *110*, 791.
- Carvalho, A. T. P.; Fernandes, P. A.; Ramos, M. J. *J. Comput. Chem.* **2006**, *27*, 966.
- Ohno, K.; Kamiya, N.; Asakawa, N.; Inoue, Y.; Sakurai, M. *Chem. Phys. Lett.* **2001**, *341*, 387.
- Li, H.; Robertson, A. D.; Jensen, J. H. *Proteins: Struct., Funct., Bioinf.* **2005**, *61*, 704.
- Mehler, E. L.; Guarnieri, F. *Biophys. J.* **1999**, *77*, 3.
- Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2006**, *110*, 10478.
- Vessecchi, R.; Galembeck, S. E. *J. Phys. Chem. A* **2008**, *112*, 4060.
- Sousa, S. F.; Fernandes, P. A.; Ramos, M. J. *J. Phys. Chem. A* **2007**, *111*, 10439.
- Zhao, Y.; Gonzalez-Garcia, N.; Truhlar, D. G. *J. Phys. Chem. A* **2005**, *109*, 2012.
- Zhao, Y.; Truhlar, D. G. *Acc. Chem. Res.* **2008**, *41*, 157.
- Goerigk, L.; Grimme, S. *J. Chem. Theory Comput.* **2011**, *7*, 291.
- Goerigk, L.; Grimme, S. *J. Chem. Theory Comput.* **2010**, *6*, 107.
- Zhao, Y.; Truhlar, D. G. *J. Chem. Theory Comput.* **2011**, *7*, 669.
- Miller, G. E.; Banerjee, N. C.; Stowe, C. M. *J. Pharmacol. Exp. Ther.* **1967**, *157*, 245-8.
- Butler, J.; Land, E. J.; Prutz, W. A.; Swallow, A. J. *J. Chem. Soc., Chem. Commun.* **1986**, 348.
- Ratilla, E. M. A.; Scott, B. K.; Moxness, M. S.; Kostic, N. M. *Inorg. Chem.* **1990**, *29*, 918.

- (24) Schaefer, J. R.; Newman, M. S.; Verhoek, F. H. *J. Am. Chem. Soc.* **1944**, *66*, 1847.
- (25) Kristol, D. S.; Krauthaim, P.; Stanley, S.; Parker, R. C. *Bioorg. Chem.* **1975**, *4*, 299.
- (26) Bruice, T. C.; Schmir, G. L. *J. Am. Chem. Soc.* **1957**, *79*, 1663.
- (27) Barone, V.; Cossi, M. *J. Phys. Chem. A* **1998**, *102*, 1995.
- (28) Cossi, M.; Rega, N.; Scalmani, G.; Barone, V. *J. Comput. Chem.* **2003**, *24*, 669.
- (29) Dunning, T. H.; Peterson, K. A.; Wilson, A. K. *J. Chem. Phys.* **2001**, *114*, 9244.
- (30) Wilson, A. K.; Dunning, T. H. *J. Chem. Phys.* **1997**, *106*, 8718.
- (31) Truhlar, D. G. *J. Chem. Phys. Lett.* **1998**, *294*, 45.
- (32) Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2005**, *109*, 6624.
- (33) Fast, P. L.; Sanchez, M. L.; Truhlar, D. G. *J. Chem. Phys.* **1999**, *111*, 2921.
- (34) Lee, J. S. *Phys. Rev. A* **2003**, 68.
- (35) Schwartz, C. *Phys. Rev.* **1962**, *126*, 1015–.
- (36) Helgaker, T.; Klopper, W.; Koch, H.; Noga, J. *J. Chem. Phys.* **1997**, *106*, 9639.
- (37) Halkier, A.; Helgaker, T.; Jorgensen, P.; Klopper, W.; Koch, H.; Olsen, J.; Wilson, A. K. *J. Chem. Phys. Lett.* **1998**, *286*, 243.
- (38) Halkier, A.; Helgaker, T.; Jorgensen, P.; Klopper, W.; Olsen, J. *J. Chem. Phys. Lett.* **1999**, *302*, 437.
- (39) Jurecka, P.; Hobza, P. *J. Chem. Phys. Lett.* **2002**, *365*, 89.
- (40) Ribeiro, A. J. M.; Ramos, M. J.; Fernandes, P. A. *J. Chem. Theory Comput.* **2010**, *6*, 2281.
- (41) 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, C. Y.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03*; Gaussian, Inc.: Wallingford, CT, 2004.
- (42) 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.
- (43) Wheeler, S. E.; Houk, K. N. *J. Chem. Theory Comput.* **2010**, *6*, 395.
- (44) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. *J. Chem. Phys.* **2010**, *132*, 154104.
- (45) AK Grimme Homepage. <http://www.uni-muenster.de/Chemie.oc/grimme/> (accessed October 2011).
- (46) Becke, A. D.; Johnson, E. R. *J. Chem. Phys.* **2005**, *123*, 154101.
- (47) Johnson, E. R.; Becke, A. D. *J. Chem. Phys.* **2005**, *123*, 024101.
- (48) Johnson, E. R.; Becke, A. D. *J. Chem. Phys.* **2006**, *124*, 174104.
- (49) Adamo, C.; Barone, V. *J. Chem. Phys.* **1998**, *108*, 664.
- (50) Becke, A. D. *J. Chem. Phys.* **1996**, *104*, 1040.
- (51) Perdew, J. P. *Phys. Rev. B* **1986**, *33*, 8822.
- (52) Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2004**, *108*, 6908.
- (53) Zhang, I. Y.; Luo, Y.; Xu, X. *J. Chem. Phys.* **2010**, *132*, 104105.
- (54) Zhang, Y.; Xu, X.; Goddard, W. A. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 4963.
- (55) Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098.
- (56) Boese, A. D.; Martin, J. M. L. *J. Chem. Phys.* **2004**, *121*, 3405.
- (57) Zhao, Y.; Truhlar, D. G. *Theor. Chem. Acc.* **2008**, *120*, 215.
- (58) Adamo, C.; Barone, V. *J. Chem. Phys. Lett.* **1997**, *274*, 242.
- (59) Lee, C. T.; Yang, W. T.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785.
- (60) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865.
- (61) Yanai, T.; Tew, D. P.; Handy, N. C. *J. Chem. Phys. Lett.* **2004**, *393*, 51.
- (62) Hamprecht, F. A.; Cohen, A. J.; Tozer, D. J.; Handy, N. C. *J. Chem. Phys.* **1998**, *109*, 6264.
- (63) Zhao, Y.; Truhlar, D. G. *J. Chem. Theory Comput.* **2005**, *1*, 415.
- (64) Krieger, J. B.; Chen, J. Q.; Iafate, G. J.; Savin, A. In *Electron Correlations and Materials Properties*; Gonis, A., Kioussis, N., Eds.; Plenum: New York, 1999; p 463.
- (65) Schmider, H. L.; Becke, A. D. *J. Chem. Phys.* **1998**, *108*, 9624.
- (66) Zhao, Y.; Truhlar, D. G. *J. Chem. Phys.* **2006**, *125*.
- (67) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648.
- (68) Grimme, S. *J. Comput. Chem.* **2006**, *27*, 1787.
- (69) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. *J. Phys. Chem.* **1994**, *98*, 11623.
- (70) Perdew, J. P. Unified Theory of Exchange and Correlation Beyond the Local Density Approximation. In *Electronic Structure of Solids '91*; Ziesche, P., Eschig, H., Eds.; Akademie Verlag: Berlin, 1991; p 11.
- (71) Tao, J. M.; Perdew, J. P.; Staroverov, V. N.; Scuseria, G. E. *Phys. Rev. Lett.* **2003**, *91*, 146401/1–4.
- (72) Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. Chem. Phys.* **2005**, *7*, 2701.
- (73) Handy, N. C.; Cohen, A. J. *Mol. Phys.* **2001**, *99*, 403.
- (74) Hoe, W. M.; Cohen, A. J.; Handy, N. C. *J. Chem. Phys. Lett.* **2001**, *341*, 319.
- (75) Lynch, B. J.; Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2003**, *107*, 1384.
- (76) Van Voorhis, T.; Scuseria, G. E. *J. Chem. Phys.* **1998**, *109*, 400.
- (77) Gill, P. M. W. *Mol. Phys.* **1996**, *89*, 433.
- (78) Boese, A. D.; Martin, J. M. L.; Handy, N. C. *J. Chem. Phys.* **2003**, *119*, 3005.
- (79) Dahlke, E. E.; Truhlar, D. G. *J. Phys. Chem. B* **2005**, *109*, 15677.
- (80) Zhao, Y.; Lynch, B. J.; Truhlar, D. G. *J. Phys. Chem. A* **2004**, *108*, 2715.
- (81) Chai, J. D.; Head-Gordon, M. *J. Phys. Chem. Chem. Phys.* **2008**, *10*, 6615.
- (82) Chai, J. D.; Head-Gordon, M. *J. Chem. Phys.* **2008**, *128*.
- (83) Grimme, S. *J. Chem. Phys.* **2006**, *124*, 034108.
- (84) Schwabe, T.; Grimme, S. *J. Phys. Chem. Chem. Phys.* **2006**, *8*, 4398.
- (85) Kormos, B. L.; Cramer, C. J. *J. Phys. Org. Chem.* **2002**, *15*, 712.
- (86) Lynch, B. J.; Fast, P. L.; Harris, M.; Truhlar, D. G. *J. Phys. Chem. A* **2000**, *104*, 4811.
- (87) Vosko, S. H.; Wilk, L.; Nusair, M. *Can. J. Phys.* **1980**, *58*, 1200.
- (88) Slater, J. C. *The Self-Consistent Field for Molecules and Solids*; McGraw-Hill: New York, 1974; Vol. 4.
- (89) Wei, Y.; Singer, T.; Mayr, H.; Sastry, G. N.; Zipse, H. *J. Comput. Chem.* **2008**, *29*, 291.
- (90) Bras, N. F.; Goncalves, R.; Fernandes, P. A.; Mateus, N.; Ramos, M. J.; de Freitas, V. *Biochemistry* **2010**, *49*, 5097.
- (91) Moon, A. F.; Edavettal, S. C.; Krahn, J. M.; Munoz, E. M.; Negishi, M.; Linhardt, R. J.; Liu, J.; Pedersen, L. C. *J. Biol. Chem.* **2004**, *279*, 45185.


Validation of Relativistic DFT Approaches to the Calculation of NMR Chemical Shifts in Square-Planar Pt²⁺ and Au³⁺ Complexes

Tomasz Pawlak,^{†,§,||} Markéta L. Munzarová,^{*,†,‡} Leszek Pazderski,^{||} and Radek Marek^{*,†,‡}

[†]National Center for Biomolecular Research, Faculty of Science and [‡]Central European Institute of Technology (CEITEC), Masaryk University, Kamenice 5/A4, CZ-62500 Brno, Czech Republic

[§]Center of Molecular and Macromolecular Studies, Polish Academy of Sciences, Sienkiewicza 112, PL-90363 Łódź, Poland

^{||}Faculty of Chemistry, Nicholas Copernicus University, Gagarina 7, PL-87100 Toruń, Poland

 Supporting Information

ABSTRACT: Recently implemented hybrid density functional methods of calculating nuclear magnetic shielding using the two-component zeroth-order regular approximation approach (*J. Phys. Chem. A* **2009**, *113*, 11495) have been employed for a series of compounds containing heavy transition-metal atoms. These include Pt²⁺, Pd²⁺, and Au³⁺ organometallics and metal complexes with azines, some of which exhibit interesting biological and catalytic activities. In this study we investigate the effects of geometry, exchange–correlation functional, solvent, and scalar relativistic and spin–orbit corrections on the nuclear magnetic shielding—mainly for ¹³C and ¹⁵N atoms connected to a heavy-atom center. Our calculations demonstrate that the B3LYP method using effective core potentials and a cc-pwCVTZ-PP/6-31G** basis set augmented with the polarizable continuum model of the dimethylsulfoxide solvent provides geometries for the complexes in question which are compatible with the experimental NMR results in terms of both the trends and the absolute values of the ¹³C shifts. The important role of the exact exchange admixture parameter for hybrid functionals based on B3LYP and PBE0 is investigated systematically for selected Pt²⁺ and Au³⁺ complexes. The ¹³C and ¹⁵N NMR chemical shifts are found to be best reproduced by using a B3LYP or PBE0 approach with 30% and 40–50% exact exchange admixtures for the Pt²⁺ and Au³⁺ complexes, respectively. The spin–orbit contributions to the ¹⁵N NMR chemical shifts reflect metal–ligand bonding that is much more ionic for the Au³⁺ than for the Pt²⁺ complex. Finally, an optimized density functional method is applied to a series of transition-metal complexes to estimate the scope and the limitations of the current approach.

1. INTRODUCTION

Azine ligands are known to coordinate with transition-metal ions, such as Pd²⁺, Pt²⁺, and Au³⁺. Many of these coordination compounds exhibit noticeable biological activities or catalytic properties.^{1–3} Cytotoxicity in particular has been reported for neutral Pd²⁺ and Pt²⁺ chloride compounds (e.g., [Pt(2ppy*)(DMSO)Cl], **1**; [Pd(2ppy*)NH₃Cl]),^{4–6} for their bromide and iodide analogs, for cations of the general structure [Pt(2ppy*)(LL)]⁺,⁷ and for [Au(2ppy*)Cl₂], **2**,^{8,9} and its analogs,^{10,11} wherein 2ppy* denotes 2-phenylpyridine (2ppy) that has been deprotonated at C-2' and acts as a chelating ligand (see Chart 1). In addition, [Au(2ppy*)Cl₂]¹² and some other related compounds^{13–16} have been used as catalysts in organic syntheses.

Many of these coordination compounds have been identified as efficient intercalators of nucleic acids.¹ In order to find the structure–activity relationships of these compounds, it is necessary to characterize their structures in detail.¹ Various experimental techniques, e.g., X-ray diffraction, IR and UV–vis spectroscopy, Raman spectroscopy, and mass spectrometry, are frequently employed for structural studies.

Nuclear magnetic resonance (NMR) spectroscopy represents one of the most powerful techniques for characterizing molecular and supramolecular structures both in solution and in the solid state.¹⁷ The NMR chemical shift is very sensitive and is probably the most widely employed NMR probe. Its isotropic value, typically measured in solution, as well as the principal components of the

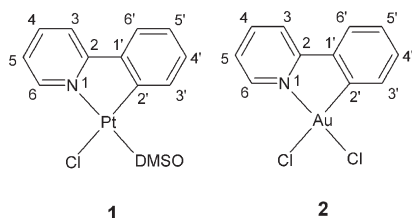
chemical-shift tensor determined for solid samples is derived from the distribution of the electron density surrounding the investigated nucleus. In principle, several nuclei can be employed in NMR studies of transition-metal complexes with heterocyclic ligands.¹⁸ The relative importance of individual effects (e.g., substitution, environment) on a particular chemical shift can be decoded by means of quantum chemical methods.¹⁹

Density functional theory (DFT) currently represents an efficient compromise between accuracy and the requirements of the computer for calculating NMR chemical shifts. It is frequently used for calculations involving small to medium size molecules or molecular assemblies. However, several physical effects must be taken into account. First, compounds containing transition metals must be treated using methods that cover electron correlation. Second, if the aim is to reproduce or predict the NMR data of condensed phases, solvent or crystal effects must be included.²⁰ Third, the choice of an appropriate reference compound is crucially important. Typically, secondary reference compounds are selected in order to eliminate the systematic errors arising due to the limitations of the currently used density functionals. Last, but not least, a theoretical description of the structures and spectral properties of compounds containing

Received: May 31, 2011

Published: November 08, 2011

Chart 1. Structures and Atom Numbering for Compound 1, *trans*(*S,N*)-[Pt(2-ppy*)(DMSO)Cl] and Compound 2, [Au(2ppy*)Cl₂]



heavy elements requires the inclusion of relativistic effects;²¹ this is particularly important for NMR parameters.^{22,23}

The most exact approach to calculating the NMR parameters of transition-metal complexes is a fully relativistic four-component theory, as implemented, for example, in the ReSpect-MAG²⁴ or Dirac²⁵ codes. Very recently, four-component codes have also been employed for calculations involving transition-metal complexes.²⁶ However, for validating various density functional approaches, the selection of the functionals within the four-component codes is rather limited. Other potentially applicable methods include the effective core potential (ECP) treatment²⁷ of scalar relativistic (SR) effects as well as the one-component (SR) or two-component inclusion of the SR and spin-orbit (SO) effects within the zeroth-order regular approximation (ZORA).^{28–33} Both approaches are employed in this study.

In the present paper, we perform a systematic validation study of DFT approaches to the calculation of the NMR chemical shifts for compounds **1** and **2** (Chart 1). Compound **1** is a well-known platinum complex with square-planar coordination,^{34–36} which has been characterized by using single-crystal X-ray diffraction.³⁷ Compound **2** is a structurally related Au³⁺ complex. We began our study with a series of geometry optimizations for molecule **1**. For the resulting structure, we tested several approaches to the calculation of NMR chemical shifts and selected one on the basis of its performance relative to experiment, comparing both the trends and the absolute values of the NMR chemical shifts. The selected NMR approach was then applied to the structures resulting from the various geometry optimizations and the X-ray diffraction results. The trends and the absolute values of the chemical shifts were again checked against the experimental data, and the suitability of the selected geometry was confirmed. Our study was completed with the application of the method to a set of four additional Pt²⁺ and Pd²⁺ square-planar complexes.

2. METHODOLOGICAL DETAILS

The quantum chemical calculations were carried out using the Gaussian03 (G03)³⁸ and Amsterdam Density Functional 2009 (ADF)³⁹ codes. For the G03 calculations, the scf = tight keyword was used (convergence to 10^{−8} in the energy and to 10^{−6} in the density matrix). An ultrafine grid with 75 radial shells and 302 angular points was employed. The geometry was always optimized without any symmetry restrictions, but all of the resultant structures possessed a horizontal plane of symmetry (group C_s).

2.1. Experimental Molecular Geometry. The experimental X-ray geometries for compound **1** (Cambridge Crystallographic Data Center code JISPAD01),^{35,40} compound **2** (CCDC code IJAQEP),⁹ compound **3** (CCDC code SOXZUA),⁴¹ and compound **5** (CCDC code BPYCPT)⁴² were obtained from the CCDC.⁴³ The geometry for compound **1** was used to calibrate

the methods employed to optimize the molecular geometry (Sections 2.2 and 3.1), where the comparisons are carried out only for the positions of atoms other than hydrogen. The data deposited at the CCDC include the positions of the hydrogens, but the resolution is insufficient for our purposes. Hence, for the purpose of computing chemical shifts, “X-ray geometry” denotes the X-ray coordinates of the atoms other than hydrogen augmented by the coordinates of the hydrogens as optimized by using B3LYP/6-31G**.

2.2. Calculated Molecular Geometry. The starting structures for the optimization of the geometry were the X-ray geometries for compounds **1–3** and **5**. The geometries of compounds **4** and **6** were calculated de novo. The following theoretical methods were used to optimize the molecular geometry: Hartree–Fock (HF),^{44,45} Møller–Plesset perturbation theory to the second order (MP2),^{46,47} and three different combinations of DFT exchange and correlation potentials, abbreviated as BLYP, B3LYP, and BHandHLYP. The functionals combine the LYP generalized gradient approximation (GGA) for correlation⁴⁸ with three different approximations for exchange, namely the Becke’s GGA for exchange (B),⁴⁹ the Becke’s three-parameter hybrid functional (B3, this includes 20% of exact exchange),⁵⁰ and the “half-and-half” hybrid incorporating as much as 50% of exact exchange.⁵¹ The aforementioned methods were chosen for the following reasons: The B3LYP method has often been reported to perform favorably in comparison with experimental and CCSD(T) structural data.⁵² The BLYP and BHandHLYP methods have been used to see the impact of the inclusion of exact exchange on the structural data. The HF method has been included as the most extreme method on the scale BLYP→B3LYP→BHandHLYP→HF. The MP2 method was a natural choice because it usually agrees well with the experimental data for the geometries of 4d and 5d species provided that relativistic effects are taken into account.⁵²

The relativistic effective core potentials (ECPs) and basis sets employed for platinum were: ECP60MDF with a 7s,7p,6d,3f,2g correlation consistent orbital basis,⁵³ ECP60MWB with a 6s,5p,3d,2f,1g orbital basis,^{54–56} and Lanl2dz with a 3s,3p,2d orbital basis.^{57–59} In the two abbreviations of the form ECP_nX_Y, *n* is the number of core electrons which are replaced by the pseudopotential, X denotes the reference system used to generate the pseudopotential (X = M: neutral atom), and Y stands for the theoretical level of the reference data (Y = Dirac–Fock (DF): relativistic, Y = Wood–Boring (WB): quasirelativistic). Lanl2dz is an acronym that stands for Los Alamos National Laboratory 2 double- ζ (density functional theory). The motive for selecting the ECPs was the generally recognized good performance of the Stuttgart–Cologne MDF and MWB pseudopotentials. The older Lanl2dz pseudopotential was tested because it is very often used in calculations that augment experimental NMR data. The orbital basis sets employed are those recommended for use with the ECPs.

The relativistic ECP ECP60MDF was employed for gold as well, while ECP28MDF (with a correlation consistent cc-pwCVTZ-PP orbital basis) was used for palladium. For nonmetal atoms we tested the Pople bases 6-31G*, 6-31G**, and 6-311G* (only the 6-31G** data are shown below).^{60–62} The nonmetal basis sets were chosen based on our previous experience with the influence of the size of the basis set on the geometrical parameters. The resulting molecular structures were reoptimized using the polarizable continuum model (PCM) of the solvent. Dimethylsulfoxide (DMSO) was used as the solvent with the following standard parameters: a dielectric constant ϵ of 46.7, a molecular

volume of 70.94 \AA^3 , a numerical density of 0.00849 \AA^{-3} , and a radius of the solvent molecule of 2.455 \AA .^{63,64} Each stationary point was characterized by calculating the harmonic vibration frequencies.

To evaluate the effects of crystal packing on the molecular geometry, we performed a B3LYP optimization (ECP60MDF/6-31G**) of a single molecule of compound **1** surrounded by eight neighboring molecules adopting a topological arrangement that has been characterized by using X-ray diffraction (see Supporting Information). In other words, only the central molecule of this cluster of nine molecules was relaxed.

2.3. Experimental NMR Chemical Shifts. The ^1H , ^{13}C , and ^{15}N NMR chemical shifts discussed in this contribution were measured in dimethylsulfoxide (DMSO- d_6) solution. The ^1H and ^{15}N NMR chemical shifts had been reported previously (see footnotes to the tables). The ^{13}C NMR data were measured as part of this work and were assigned by using the standard 2D ^1H – ^{13}C NMR correlation techniques, gradient-selected heteronuclear single-quantum correlation (gs-HSQC)⁶⁵ and gradient-selected heteronuclear multiple-bond correlation (gs-HMBC).^{66,67} The ^1H and ^{13}C NMR chemical shifts are reported relative to tetramethylsilane and the ^{15}N NMR chemical shifts relative to liquid ammonia.^{68,69}

2.4. Nuclear Magnetic Shielding—G03. The selected geometry was used to calculate the nuclear magnetic shielding by means of DFT and ab initio (HF and MP2) methods using the gauge including atomic orbitals (GIAO) approach.^{70–72} We tested the BLYP, B3LYP, and BHandHLYP functionals previously employed to optimize the geometry. We also tested three further combinations of exchange and correlation potentials, abbreviated as BP86, PW91, and PBE0. The first combines Becke's GGA for exchange⁴⁹ with the P86^{73,74} GGA for correlation. The second uses both the exchange and the correlation parts of the PW91 functional.^{75–77} The third is the hybrid functional of Perdew, Burke, and Ernzenhof.^{78,79} Two basis sets were used for platinum: cc-pwCVTZ-PP (with relativistic ECP ECP60MDF) and Fægri-III^{80,81} (FIII) based on the primitive exponents of Fægri⁸² and contracted as well as polarized similarly to the IGLO-III basis set.⁸³ An optimized basis for light atoms (IGLO-III-UT3, cf. below) was employed. The NMR calculations were performed in vacuo (on an in vacuo optimized geometry) and also using the PCM solvent model (on a PCM reoptimized geometry).

2.5. Nuclear Magnetic Shielding—ADF. All GIAO shielding calculations in ADF were performed using the polarized valence triple- ζ (TZP) basis set from the standard ADF library along with, first, the local density approximation, employing the exchange potential⁸⁴ combined with the correlation potentials due to Perdew and Wang (PW92)⁸⁵ and to Vosko, Wilk, and Nusair (VWN).⁸⁶ Second, we employed the GGA functionals BP86, PW91, and BLYP described above and Keal–Tozer functionals KT1 and KT2⁸⁷ which had been optimized for the calculation of chemical shifts. Finally, the hybrid functionals B3LYP, BHandHLYP, and PBE0 were employed. For the functionals B3LYP and PBE0, we systematically investigated the influence of the exact exchange admixture (0–50%) on the nuclear shielding. The latter study was done at the nonrelativistic level using the conductor-like screening (COSMO)^{88,89} approach for the DMSO solvent as implemented in ADF⁹⁰ and at the ZORA-SO relativistic level using COSMO.

2.6. Calculated NMR Chemical Shifts. The choice of the reference compound can help to reduce systematic errors in the DFT calculation of chemical shifts. In this study, we selected

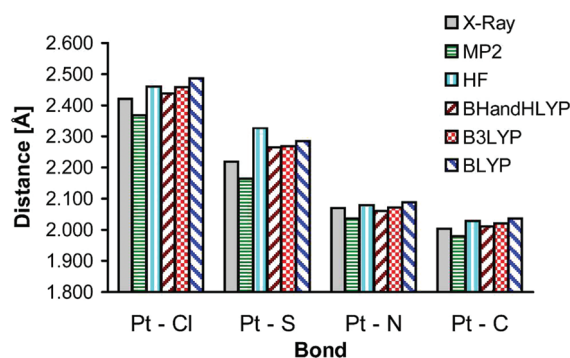


Figure 1. Interatomic distances from the platinum center for compound **1**, optimized using different functionals with ECP60MDF/cc-pwCVTZ-PP for Pt and the 6-31G** basis set for light atoms.

benzene as the secondary reference⁹¹ for ^1H and ^{13}C NMR (residual benzene- d_5 in benzene- d_6 for ^1H NMR: $\delta = 7.15$ ppm, benzene- d_6 for ^{13}C NMR: $\delta = 127.8$ ppm), and pyridine in DMSO- d_6 as the secondary reference for ^{15}N NMR ($\delta = 316.9$ ppm relative to liquid ammonia).^{68,69,92} The chemical shifts δ_i were obtained using the equation:

$$\delta_i = \sigma_{\text{ref}} - \sigma_i + \sigma_{\text{ref}} \quad (1)$$

where σ_{ref} is the calculated shielding constant of a given nucleus in the reference compound, σ_i is the calculated shielding constant of the investigated nucleus, and δ_{ref} is the experimental chemical shift of the secondary reference relative to the primary standard.

2.7. SR and SO Contributions to the Chemical Shift. The SR contribution to the nuclear shielding was calculated as the difference between the SR-ZORA calculation and the fully non-relativistic calculation in the ADF program. The SR contribution to the chemical shift (Δ_{SR}) was then calculated as

$$\Delta_{\text{SR}} = \sigma_{\text{SR,ref}} - \sigma_{\text{SR},i} \quad (2)$$

where $\sigma_{\text{SR,ref}}$ is the SR contribution to the nuclear shielding constant in the reference compound and $\sigma_{\text{SR},i}$ is the SR contribution to the nuclear shielding constant of the investigated nucleus. Similarly, the SO contribution to the chemical shift was calculated as

$$\Delta_{\text{SO}} = \sigma_{\text{SO,ref}} - \sigma_{\text{SO},i} \quad (3)$$

where $\sigma_{\text{SO,ref}}$ is the SO contribution to the nuclear shielding constant in the reference compound and $\sigma_{\text{SO},i}$ is the SO contribution to the nuclear shielding constant of the investigated nucleus.

2.8. Visualizations. The highest molecular orbitals (MOs) relevant for the Δ_{SO} contribution are visualized in Figures 10 and 11 by using the Molekel program.⁹³

3. RESULTS AND DISCUSSION

3.1. Geometries of Compounds 1 and 2. A series of ab initio and DFT geometry optimizations were performed for compound **1** as characterized by two well-resolved X-ray structures in order to find a suitable computational method for optimizing the structures of complexes **2–5**. In Figure 1, the resulting geometries are compared to the JISPAD01 structure (low-temperature structure, more symmetric crystal, lower value of R_{int} and lower residual electron density as compared to the JISPAD structure) in terms of the metal–ligand bond lengths.

The comparison of gas-phase optimized structures and the X-ray geometries obtained for a crystal may appear questionable since crystal-packing effects can, in principle, influence bond lengths substantially. In order to estimate the influence of the crystal packing on the geometry of complex **1**, we performed a B3LYP optimization (with the same parameters as for the other optimizations) of a single molecule of **1** surrounded by its eight nearest neighbors (for atomic coordinates, see Supporting Information). The starting structure was the X-ray geometry of the nine molecules, and only the central molecule of the cluster was relaxed. The resulting bond lengths are given in Table S1, Supporting Information. Compared to the gas-phase optimized counterpart, the bond lengths within the embedded molecule varied by 0.03 Å for the Pt–Cl bond, and by only 0.01 Å or even less for the Pt–S, Pt–N, and Pt–C bonds. Thus, for the Pt–S, Pt–N, and Pt–C bonds, the crystal packing effects turn out to be negligible, and the corresponding bond lengths obtained from the X-ray structures can serve as the standards for evaluating the suitability of various methods of obtaining structures. A comparative calculation showed that the effect of using the structure optimized in the crystal embedding instead of the gas-phase structure results in changes in the chemical shift of 1.3 ppm for N-1, 2.5 ppm for C-2', and 0.5 ppm or less for the remaining atoms. These numbers represent the maximum order of accuracy which can be expected for a comparison between the chemical shielding computed based on X-ray geometry and the NMR shifts obtained experimentally in solution, as presented later in this work.

The metal–carbon, metal–nitrogen, and metal–sulfur bond lengths for complex **1** are obviously best reproduced by the hybrid density functional approaches B3LYP and BHandHLYP. The distances shorten with increasing exact exchange admixture in the density functional; this is most apparent for the Pt–Cl and Pt–S bonds. In agreement with earlier findings,⁹⁴ the HF method itself, however, significantly overestimates the Pt–S bond length, while MP2 underestimates all metal–ligand distances. Of the two suitable hybrid density functionals, we have selected the well-established B3LYP approach to the geometry for the following calculations of the NMR shielding constants.

For comparative purposes, we have also performed a series of geometry optimizations employing the ZORA approach. The results, cf. Table S1, Supporting Information, reveal that for the Pt–Cl, Pt–N, and Pt–C bonds, the ZORA bond lengths are 0.01–0.02 Å longer than the ECP MDF bond lengths, while the Pt–S bond length stays approximately constant. The ZORA-SO bond lengths differ from the ZORA bond lengths by less than 0.01 ppm, i.e., the effect of the SO coupling on the geometry is negligible.

In computational studies supplementing routine NMR investigations of transition-metal complexes,^{36,95,96} the B3LYP optimization of the geometry is frequently performed with the large-core Lanl2dz effective core potential available in Gaussian. It is employed despite persuasive theoretical results that demonstrate that large-core ECPs are insufficient for both the structures and the NMR parameters and indicate that small-core and relativistic ECPs must be used instead.^{97,98} To evaluate the effect of the particular pseudopotential on the geometry, we performed a simple comparative study using the B3LYP optimizations with the small-core quasirelativistic MWB ECP and the large-core Lanl2dz ECP, using the well-established small-core relativistic MDF ECP as reference. The MWB and Lanl2dz ECPs produce Pt–N and Pt–C bond distances that are quite close to those

Table 1. Interatomic Distances (in Å) around the Platinum Center in Compound **1** Calculated on an in Vacuo Optimized Geometry Using Various Relativistic ECPs and Basis Sets

	Pt–Cl	Pt–S	Pt–N	Pt–C
MDF ^a	2.458	2.268	2.072	2.021
MWB ^b	2.469	2.277	2.074	2.020
LA ^c	2.508	2.322	2.071	2.020

^a Geometry optimized using B3LYP/ECP60MDF/cc-pwCVTZ-PP+6-31G**. ^b Geometry optimized using B3LYP/ECP60MWB/ECP60MWB+6-31G**. ^c Geometry optimized using B3LYP/Lanl2dz/Lanl2dz+6-31G**.

given by the reference MDF (Table 1). However, this is not true for the Pt–Cl and Pt–S bonds, where Lanl2dz produces distances which are 0.04–0.05 Å greater than those given by MDF. In addition, the differences between experiment and theory for these two bonds are twice as large for the Lanl2dz ECP (ca. 0.1 Å) as for the Stuttgart ECPs (ca. 0.05 Å). Although one might consider these differences for Pt–Cl and Pt–S bonds to be relatively unimportant in a ¹³C and ¹⁵N NMR study, the opposite is true due to the indirect trans-substituent effects of the chlorine and sulfur atoms on the nuclear shieldings of ¹³C and ¹⁵N.

The structure of complex **2** was optimized using the BLYP, B3LYP, BHandHLYP, HF, and MP2 methods as described in the Methodological Details Section. The resulting metal–ligand bond lengths are compared to the available X-ray data in Figure S1, Supporting Information. As for complex **1**, the MP2 bond lengths are the shortest, while, contrary to complex **1**, the BLYP bond lengths are in all cases the longest. A comparison with experiment might indicate that the BHandHLYP bond lengths are “best” for the Au–Cl bonds, while the MP2 bond lengths are “best” for the Au–C2' and Au–N bonds. However, since the resolution of the X-ray structure (CCDC code IJAQEP)³⁵ is relatively poor and a test NMR calculation on the X-ray structure gave meaningless results (cf. below), we consider the X-ray structure to be useless for comparing the bond lengths.

3.2. Nuclear Magnetic Shieldings for Compounds **1** and **2**.

Below, we perform a systematic study for compound **1** in order to identify the optimum method for calculating the NMR chemical shifts. The method is characterized by the level of calculation of the nonrelativistic nuclear magnetic shielding (we test various basis sets and density functionals there), the inclusion of the surroundings (gas-phase calculation, continuum model), and the treatment of relativistic effects (nonrelativistic, SR, or including both SR and SO effects). When comparing the computed and experimental data, we try, where possible, to decouple these individual effects by determining atoms for which some of the influences could prove negligible. Furthermore, we use the additivity of the individual effects when this can be demonstrated. Finally, we focus on trends in the nuclear magnetic shielding within the whole molecule. This enables us to judge the quality of the description globally, which is always more reliable than attempting to “fit” the individual shifts to experiment as closely as possible.

3.2.1. Basis for Nonmetal Atoms: A Convergence Study for Compound **1.** The convergence of the standard IGLO-III⁸³ basis set (also known as Huzinaga–Kutzelnig HIII)⁹⁹ for calculating the chemical shifts of the light atoms C-2' and N-1 was carefully tested at the SR level. The results are shown in Figure 2. The procedure was done independently for each of the two atoms in the following way: First, the standard IGLO-III was uncontracted

(IGLO-III-U). Second, one extra function was added to each type of angular momentum (s, p, and d) with an exponent three times as large as that already available. (The values of the high exponents have been chosen according to the fact that in variationally optimized bases, the ratio between two successive exponents is approximately constant.¹⁰⁰ Within the IGLO-III basis set, this ratio is close to 3.) The addition of the extra functions was done systematically in three steps (resulting in the bases IGLO-III-UT1, IGLO-III-UT2, and IGLO-III-UT3). Third, three additional diffuse functions were continually added to each angular momentum type with an exponent one-third as large as that already available, resulting in the bases UT3_D1, UT3_D2, and UT3_D3. Finally, one extra polarization function of *f* type was added from the cc-pVQZ basis set.¹⁰¹ As demonstrated in

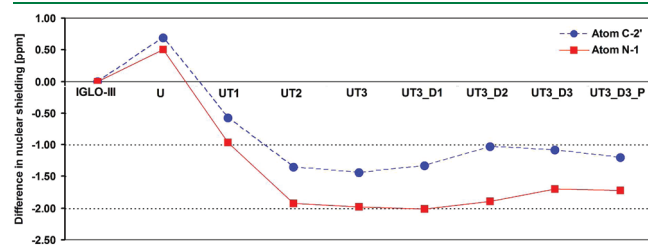


Figure 2. Convergence of the IGLO-III basis set (B3LYP with ECP60MDF/cc-pwCVTZ-PP for Pt and IGLO-III for light atoms) for atoms N-1 and C-2' in compound 1. U, uncontracted; UT1, uncontracted with one set of tight functions added, etc.; UT3_D1, uncontracted with three sets of tight functions and one set of diffuse functions, etc.; and UT3_D3_P, including one additional *f* polarization function. The differences in the nuclear magnetic shielding relative to IGLO-III are plotted on the y axis.

Figure 2, going from the IGLO-III to the IGLO-III-UT3 basis decreases the nuclear shieldings of C-2' and N-1 by ca. 1.5 and 2 ppm, respectively. Augmenting the basis further with the diffuse and polarization functions increases the nuclear shieldings by ca. 0.3 ppm. Thus, we consider the “middle-size” IGLO-III-UT3 as a very good compromise between cost and accuracy and apply it in all subsequent calculations.

For comparative purposes, we have also tested a few Jensen's pcS-*n* bases¹⁰² for the light atoms, namely pcS-2, pcS-3, and pcS-4, of which the pcS-3 basis is closest in size to our IGLO-III-UT3 basis. The nuclear magnetic shieldings obtained with pcS-3 are ca. 0.5 ppm more negative than those calculated by using the IGLO-III-UT3, whereas the pcS-2 and pcS-4 shielding constants of C-2' and N-1 lie within 1.2 ppm of those calculated by using pcS-3.

3.2.2. NR Terms and Solvent Corrections (SOLV): B3LYP Results for 1. The relatively most important solvent effects on the NMR chemical shift can be expected for the hydrogen atoms, which are located in the outer sphere of the molecule and are thus most susceptible to potential solute–solvent interactions. This is indeed evident from Table 2, where the in vacuo results, the PCM results (obtained with G03), and the COSMO results (obtained with ADF) are summarized and compared. $\Delta\delta_{\text{solv}}$ is defined as $\Delta\delta_{\text{solv}} = \delta_{\text{solv}} - \delta_{\text{in vacuo}}$ where δ_{solv} and $\delta_{\text{in vacuo}}$ denote the chemical shifts resulting from the geometry optimization and the chemical shift calculation with and without the implicit solvent model, respectively. The relative influence of the implicit solvent is on average 8% for the hydrogens but as little as 2% for the carbons and the nitrogen. Since the root-mean-square deviation (RMSD) values for the chemical shifts of the hydrogen atoms after the inclusion of solvent effects are 3–4 times smaller than those calculated in vacuo, the inclusion of the implicit solvent

Table 2. Analysis of Implicit Solvent Effects ($\Delta\delta_{\text{solv}}$ in ppm) for Compound 1^a

atom	$\delta_{\text{G03,NR}}$ (in vacuo) ^b	$\delta_{\text{G03,NR}}$ (PCM) ^c	$\Delta\delta_{\text{solv}}$	$\delta_{\text{ADF,NR}}$ (in vacuo) ^d	$\delta_{\text{ADF,NR}}$ (COSMO) ^e	$\Delta\delta_{\text{solv}}$
H-3	7.26	7.87	+0.61	7.15	7.75	+0.60
H-4	7.26	7.92	+0.66	7.19	7.83	+0.64
H-5	6.57	7.19	+0.62	6.55	7.17	+0.62
H-6	9.94	9.52	−0.42	9.97	9.43	−0.54
H-3'	8.65	8.52	−0.13	8.64	8.56	−0.08
H-4'	6.95	7.28	+0.33	6.84	7.16	+0.32
H-5'	6.76	7.21	+0.45	6.62	7.11	+0.49
H-6'	7.26	7.82	+0.56	7.11	7.75	+0.64
RMSD	0.64	0.20		0.71	0.24	
N-1	256.3	252.7	−3.6	287.2	282.1	−5.1
C-2	169.3	167.9	−1.4	169.0	166.6	−2.4
C-3	115.1	118.4	+3.3	115.5	119.3	+3.8
C-4	137.4	141.6	+4.2	137.1	141.8	+4.7
C-5	117.7	121.4	+3.7	118.6	123.1	+4.5
C-6	154.2	152.5	−1.7	152.9	151.6	−1.3
C-1'	145.0	147.2	+2.2	145.8	147.7	+1.9
C-2'	184.4	182.2	−2.2	184.1	180.4	−3.7
C-3'	136.8	134.7	−2.1	136.4	134.5	−1.9
C-4'	131.1	130.8	−0.3	130.6	130.8	+0.2
C-5'	123.2	125.3	+2.1	122.9	125.7	+2.8
C-6'	122.2	124.7	+2.5	121.5	124.8	+3.3
RMSD	13.8/3.6 ^f	12.8/1.8 ^f		13.6/3.3 ^f	12.2/1.5 ^f	

^a All chemical shifts obtained with B3LYP. ^b G03/FIII/IGLO-III-UT3//G03/ECP60MDF/cc-pwCVTZ-PP/6-31G**. ^c G03/FIII/IGLO-III-UT3/PCM//G03/ECP60MDF/cc-pwCVTZ-PP/6-31G**/PCM. ^d ADF/TZP//G03/ECP60MDF/cc-pwCVTZ-PP/6-31G**. ^e ADF/TZP/COSMO//G03/ECP60MDF/cc-pwCVTZ-PP/6-31G**/PCM. ^f RMSD for all carbons/RMSD for all carbons excluding C-2'.

Table 3. SR Contributions to the NMR Chemical Shifts (δ in ppm) for Complex **1**^a Calculated Using B3LYP

atom	$\delta_{G03,NR}^b$	$\delta_{G03,SR}^c$	$\Delta_{SR,ECP}^d$	$\delta_{ADF,NR}^d$	$\delta_{ADF,SR}^e$	$\Delta_{SR,ZORA}^e$
N-1	252.7	240.4	-12.3	255.2	244.4	-10.8
C-2	167.9	164.0	-3.9	166.6	162.6	-4.0
C-3	118.4	118.1	-0.3	119.3	119.2	-0.1
C-4	141.6	141.9	+0.3	141.8	142.2	+0.4
C-5	121.4	121.6	+0.2	123.1	123.3	+0.2
C-6	152.5	151.2	-1.3	151.6	150.5	-1.1
C-1'	147.2	143.7	-3.5	147.7	144.4	-3.3
C-2'	182.2	173.2	-9.0	180.4	171.8	-8.6
C-3'	134.7	135.4	+0.7	134.5	135.6	+1.1
C-4'	130.8	131.3	+0.5	130.8	131.4	+0.6
C-5'	125.3	124.2	-1.1	125.7	124.7	-1.0
C-6'	124.7	124.3	-0.4	124.8	124.6	-0.2
H-3	7.87	7.83	-0.04	7.75	7.80	+0.05
H-4	7.92	7.87	-0.05	7.83	7.84	+0.01
H-5	7.19	7.20	+0.01	7.17	7.21	+0.04
H-6	9.52	9.36	-0.16	9.43	9.35	-0.08
H-3'	8.52	8.43	-0.09	8.56	8.54	-0.02
H-4'	7.28	7.24	-0.04	7.16	7.17	+0.01
H-5'	7.21	7.11	-0.10	7.11	7.05	-0.06
H-6'	7.82	7.74	-0.08	7.75	7.73	-0.02

^a Calculations on a structure optimized in PCM, employing DMSO, and using B3LYP/ECP60MDF/cc-pwCVTZ-PP for Pt and 6-31G** for light atoms in the G03 program. ^b NMR chemical shifts calculated using G03/B3LYP/Fæagri-III for Pt and IGLO-III-UT3 for light atoms, PCM. ^c G03/B3LYP/ECP60MDF/cc-pwCVTZ-PP for Pt and IGLO-III-UT3 for light atoms, PCM. ^d NMR chemical shifts calculated using ADF/B3LYP/TZP/COSMO. ^e ADF/B3LYP/SR-ZORA/TZP/COSMO.

decidedly improves the model. Indeed, the implicit solvent can be well decoupled from the other theoretical parameters since the influences of the relativistic effects and the density functional on the ¹H NMR chemical shifts are much smaller in our systems (cf. Table 3 and Table S2, Supporting Information). The values calculated by using the PCM and COSMO models are very similar (differences being smaller than 0.12 ppm). We employed the COSMO model as implemented in ADF in the production calculations.

3.2.3. SR and SO Contributions: Results for Compound 1. To determine the importance of the SR effects, we calculated SR chemical shifts in G03 using relativistic ECP (ECP60MDF with the cc-pwCVTZ-PP basis set) along with several pure GGA (BP86, PW91, BLYP) and hybrid GGA (B3LYP and BHandHLYP) functionals. As an alternative, the SR chemical shifts were calculated in the ADF program using the all-electron ZORA (scalar ZORA). The SR correction to the chemical shift (Δ_{SR}) was then calculated as the difference between the SR chemical shift (δ_{SR}) and its non-relativistic counterpart (δ_{NR}):

$$\Delta_{SR} = \delta_{SR} - \delta_{NR} \quad (4)$$

where δ_{SR} is the SR chemical shift (G03_{ECP} or ADF_{ZORA}) and δ_{NR} is the nonrelativistic chemical shift calculated using G03 (FIII/IGLO-III-UT3) or ADF (TZP).

The results for both of these approaches ($\delta_{G03,SR}$ and $\delta_{ADF,SR}$) using the B3LYP functional as an example are summarized in Table 3. Values for the other functionals tested in this study are collected in Table S3, Supporting Information. The results

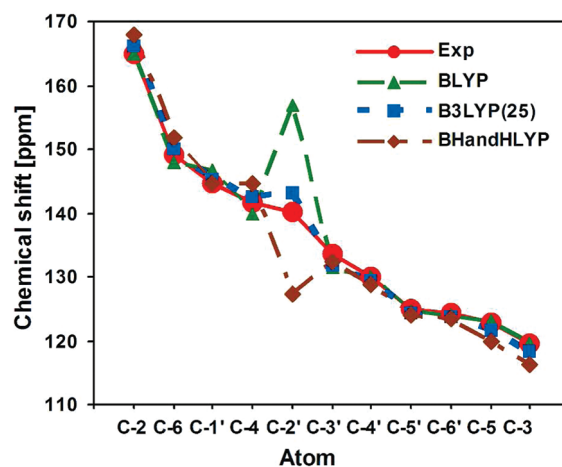


Figure 3. Experimental ¹³C NMR chemical shifts for compound **1** and those calculated by using BLYP, B3LYP(25), and BHandHLYP functionals (TZP/ZORA-SO/COSMO). A fixed B3LYP optimized structure (ECP60MDF/cc-pwCVTZ-PP for Pt and 6-31G** for light atoms, PCM) was employed for all calculations. The atoms are placed on the x axis according to decreasing experimental NMR chemical shift.

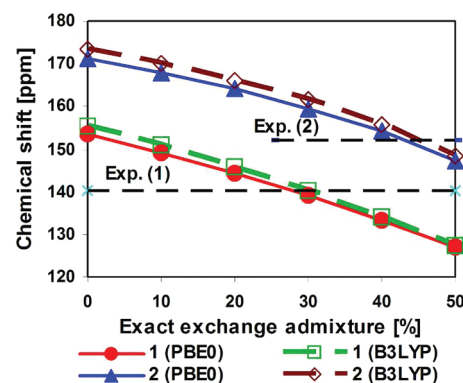


Figure 4. NMR chemical shift for C-2' in compounds **1** and **2** as a function of the exact exchange admixture for the PBE0 and B3LYP functionals calculated using the ADF/TZP/ZORA-SO/COSMO approach. Experimental values are shown as dashed lines.

demonstrate excellent agreement between the SR effects covered by relativistic ECP ($\Delta_{SR,ECP}$) and those calculated using the all-electron treatment of the SR ZORA approach ($\Delta_{SR,ZORA}$).

The second significant relativistic contribution is the SO term, which has long been recognized as important for the NMR chemical shift in compounds containing heavy atoms.^{103–106} Although the SR and SO contributions are partly self-compensating for some of the carbon and hydrogen atoms, they operate in the same direction for C-2' and N-1 and also leave non-negligible total relativistic corrections ($\Delta_{SR} + \Delta_{SO}$) for C-6, C-1', and C-3'. The influence of the density functional on all of the NR, SR, and SO terms is investigated systematically in Section 3.2.4.

3.2.4. NR, SR, and SO Terms: Influence of the Density Functional for Compounds 1 and 2. Various local density approximation and GGA methods (PW92, VWN, BP86, PW91, BLYP, KT1, KT2) give very similar NR chemical shift values (see Table S2 and Figures S2 and S3, Supporting Information). The chemical shift variation for individual functionals is within 1 ppm for all carbons except C-2' (where it reaches 2.5 ppm) and within 0.05 ppm for

all hydrogens. The effect of the exact exchange admixture is much more significant: Variations for BLYP, B3LYP, and BHandHLYP are within 7 ppm for C-2', 5 ppm for the remaining carbons, and 0.3 ppm for the hydrogens.

The inclusion of SR and SO contributions (Tables S3 and S4, Supporting Information) makes the dependence of the NMR chemical shift on the density functional much more pronounced, especially for the N and C atoms bonded directly to the transition-metal center. The NMR chemical shifts for all of the carbon atoms of compound **1** obtained by the BLYP, B3LYP (with 25% of exact exchange) and BHandHLYP functionals are compared in Figure 3.

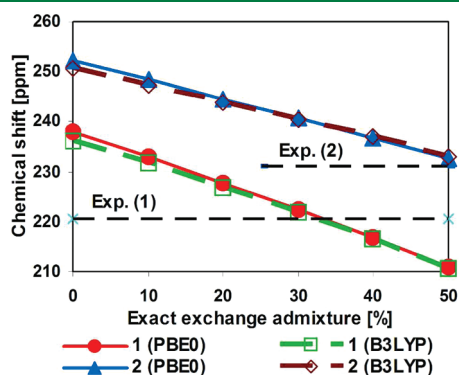


Figure 5. NMR chemical shift for N-1 in compounds **1** and **2** as a function of the exact exchange admixture for the PBE0 and B3LYP functionals calculated using the ADF/TZP/ZORA-SO/COSMO approach. Experimental values are shown as dashed lines.

The influence of 50% of exact exchange amounts to only up to 0.3 ppm for the hydrogens (i.e., less than the effect of solvation), cf. Table S5, Supporting Information, but it reaches up to 4 ppm for the carbon atoms. However, for the carbon atom C-2', which is bonded directly to the metal center, the effect is as great as 30 ppm! We note that similar dramatic effects of HF exchange have previously been reported for transition-metal NMR shifts.¹⁰⁷

The amount of the exact exchange admixture is quantitatively more important than a particular DFT exchange or correlation functional within our testing set. Indeed, the effect of the 0 to 50% exact exchange admixture on N-1 in compound **1** is as large as 32 ppm, while our ZORA-SO results for the PW92, VWN, BP86, PW91, and BLYP functionals (Table S4, Supporting Information) are within 3 ppm for all carbons including C-2' and within 5 ppm for N-1. However, it should be noted that a small dependence on a particular functional can arise due in part to the adiabatic local density approximation employed for the exchange–correlation kernel within the ADF.

Two classes of hybrid functionals have been popular in recent NMR studies: those based on the BLYP and those based on the PBE exchange–correlation functional. In order to validate their performance for our systems in detail, we studied the effect of the exact exchange admixture from 0 to 50% on a 10% grid for both classes (B3LYP and PBE0). At this point, we also introduced compound **2** in the NMR study, in order to compare the two structurally related isoelectronic species. The computed chemical shifts are plotted in Figures 4 and 5 as functions of increasing exact exchange admixture. Obviously, both the C-2' and the N-1 chemical shifts decrease as the exact exchange admixture

Table 4. NMR Chemical Shifts for Compound **1** Calculated by ADF using PBE0(30)/TZP/ZORA-SO/COSMO with Solvent Effects ($\Delta\delta_{\text{solv}}$) and SR (Δ_{SR}) and SO (Δ_{SO}) Contributions Shown Separately

atom	δ_{NR} (in vacuo) ^a	$\Delta\delta_{\text{solv}}$	$\delta_{\text{NR,solv}}$ (COSMO) ^b	Δ_{SR}	$\delta_{\text{NR,solv}} + \Delta_{\text{SR}}$ ^c	Δ_{SO}	$\delta_{\text{NR,solv}} + \Delta_{\text{SR}} + \Delta_{\text{SO}}$ ^c	expt ^e
N-1	258.5	−5.6	252.9	−12.4	240.5	−18.1	222.4	220.6
C-2	169.3	−2.6	165.1	−4.2	160.9	+3.8	164.7	165.0
C-3	116.4	+3.8	118.7	−0.5	118.2	+0.1	118.3	119.5
C-4	138.7	+4.8	142.0	+0.2	142.2	+0.5	142.7	141.6
C-5	119.6	+4.6	122.6	0.0	122.6	−0.6	122.0	122.8
C-6	154.3	−1.2	151.6	−2.1	149.5	+0.7	150.2	149.2
C-1'	145.6	+1.7	145.8	−3.6	142.2	+1.8	144.0	144.5
C-2'	181.5	−4.2	175.8	−9.0	166.8	−27.7	139.1	140.2
C-3'	137.4	−1.8	134.1	+0.7	134.8	−3.1	131.7	133.5
C-4'	131.6	+0.5	130.6	+0.3	130.9	−1.2	129.7	130.0
C-5'	124.0	+3.0	125.4	−1.3	124.1	+0.4	124.5	124.9
C-6'	122.8	+3.4	124.7	−0.6	124.1	−0.3	123.8	124.3
RMSD	12.8/3.1 ^d		10.8/1.0 ^d		8.2/1.7 ^d		0.9/0.9 ^d	
H-3	7.15	+0.63	7.78	+0.08	7.86	−0.09	7.77	8.14
H-4	7.22	+0.68	7.90	+0.04	7.94	+0.01	7.95	8.12
H-5	6.55	+0.66	7.21	+0.07	7.28	−0.07	7.21	7.50
H-6	9.79	−0.41	9.38	−0.07	9.31	+0.08	9.39	9.48
H-3'	8.59	−0.07	8.52	0.00	8.52	−0.02	8.50	8.21
H-4'	6.83	+0.36	7.19	+0.03	7.22	−0.13	7.09	7.13
H-5'	6.62	+0.53	7.15	−0.03	7.12	0.00	7.12	7.17
H-6'	7.13	+0.67	7.80	0.00	7.80	−0.15	7.65	7.77
RMSD	0.68		0.22		0.19		0.21	

^a ADF/PBE0(30)/TZP//G03/B3LYP/ECP60MDF/cc-pwCVTZ-PP+6-31G**.

^b ADF/PBE0(30)/TZP/COSMO//G03/B3LYP/ECP60MDF/cc-pwCVTZ-PP+6-31G**/PCM.

^c ADF/PBE0(30)/TZP/ZORA-SO/COSMO//G03/B3LYP/ECP60MDF/cc-pwCVTZ-PP+6-31G**/PCM.

^d RMSD for all carbons/RMSD for all carbons excluding C-2'.

^e Data from reference Pazderski, L.; Pawlak, T.; Sitkowski, J.; Kozerski, L.; Szlyk, E. *Magn. Reson. Chem.* **2009**, *47*, 932.

Table 5. NMR Chemical Shifts for Compound 2 Calculated by ADF using PBE0(40)/TZP/ZORA-SO/COSMO with Solvent Effects ($\Delta\delta_{\text{solv}}$) and SR (Δ_{SR}) and SO (Δ_{SO}) Contributions Shown Separately

atom	δ_{NR} (in vacuo) ^a	$\Delta\delta_{\text{solv}}$	$\delta_{\text{NR,solv}}$ (COSMO) ^b	Δ_{SR}	$\delta_{\text{NR,solv}} + \Delta_{\text{SR}}$ ^c	Δ_{SO}	$\delta_{\text{NR,solv}} + \Delta_{\text{SR}} + \Delta_{\text{SO}}$ ^c	expt ^e
N-1	258.7	-0.6	258.1	-21.1	237.0	-0.2	236.8	231.1
C-2	168.9	-1.9	167.0	-4.8	162.2	+2.9	165.1	163.9
C-3	116.0	+5.1	121.1	-0.9	120.2	0.0	120.2	122.2
C-4	140.3	+5.2	145.5	+0.4	145.9	+0.3	146.2	143.9
C-5	119.8	+4.3	124.1	+0.1	124.2	-0.6	123.6	125.3
C-6	153.6	-1.7	151.9	-1.6	150.3	-0.4	149.9	148.0
C-1'	143.6	+2.2	145.8	-4.1	141.7	-0.1	141.6	142.9
C-2'	188.4	+9.2	197.6	-27.9	169.7	-15.6	154.1	152.2
C-3'	137.4	-2.8	134.6	-1.3	133.3	-3.1	130.2	130.0
C-4'	132.1	+1.1	133.2	+0.5	133.7	-1.8	131.9	131.6
C-5'	127.7	+2.2	129.9	-1.9	128.0	+0.7	128.7	129.3
C-6'	122.6	+5.4	128.0	-1.1	126.9	-0.7	126.2	126.7
RMSD	11.8/4.6 ^d		13.9/2.5 ^d		5.6/1.9 ^d		1.5/1.4 ^d	
H-3	7.24	+0.61	7.85	+0.10	7.95	-0.08	7.87	8.41
H-4	7.42	+0.67	8.09	+0.03	8.12	+0.01	8.13	8.38
H-5	6.81	+0.57	7.38	+0.08	7.46	-0.12	7.34	7.76
H-6	10.21	-0.85	9.36	-0.15	9.21	+0.12	9.33	9.52
H-3'	8.50	-0.33	8.17	-0.10	8.07	-0.34	7.73	7.81
H-4'	6.90	+0.35	7.25	+0.09	7.34	-0.20	7.14	7.36
H-5'	6.83	+0.55	7.38	-0.02	7.36	-0.02	7.34	7.47
H-6'	7.25	+0.66	7.91	0.00	7.91	-0.26	7.65	7.97
RMSD	0.81		0.30		0.26		0.30	

^a ADF/PBE0(40)/TZP//G03/B3LYP/ECP60MDF/cc-pwCVTZ-PP+6-31G**. ^b ADF/PBE0(40)/TZP/COSMO//G03/B3LYP/ECP60MDF/cc-pwCVTZ-PP+6-31G**/PCM. ^c ADF/PBE0(40)/TZP/ZORA-SO/COSMO//G03/B3LYP/ECP60MDF/cc-pwCVTZ-PP+6-31G**/PCM. ^d RMSD for all carbons/RMSD for all carbons excluding C-2'. ^e Data from reference Pazderski, L.; Pawlak, T.; Sitkowski, J.; Kozerski, L.; Szlyk, E. *Magn. Reson. Chem.* **2009**, *47*, 932.

increases. Results closest to the experimental data were obtained by the PBE0 functional with 30% of exact exchange, denoted later as PBE0(30) for complex 1 and with the PBE0 functional with 40–50% of exact exchange, denoted later as PBE0(40) and PBE0(50), for complex 2. The optimum amount of the exact exchange in the density functional thus turns out to be different for the complexes of Pt and Au. This is by no means a unique example of such an observation.¹⁰⁸

Considering the NMR chemical shifts of all of the atoms, results slightly closer to the experimental values were obtained by using PBE0(30) for 1 (Table 4) and PBE0(40) for 2 (Table 5) as compared to the B3LYP(30) results for 1 (Table S5, Supporting Information) and the B3LYP(40) results for 2 (Table S6, Supporting Information). Here, the number in parentheses stands for the percentage of exact exchange substituted for the DFT exchange. Consequently, the PBE0 functional with 30 or 40% of exact exchange admixture has been employed for comparative purposes on a larger set of compounds, including the Pd²⁺ species.

Analysis of the individual NR, SR, and SO contributions to the NMR chemical shift of carbon C-2' in compound 1 reveals that the SR contribution remains practically constant for the various density functionals (see Figure 6). The NR part varies slightly, and the SO part is mainly responsible for any changes in the total chemical shift. The dependence observed for N-1 is similar to that for C-2', with the SR variability slightly increased (Figure 7). These results are fully compatible with those obtained from a SR study and a fully NR study of the influence of the density functional (Tables S2–S4, Supporting Information).

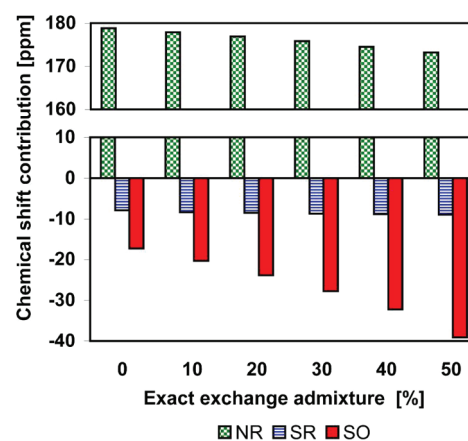


Figure 6. NR, SR, and SO contributions to the NMR chemical shift for C-2' in compound 1 calculated by using a PBE0 functional with various amounts of exact exchange.

Interestingly, for compound 2, the individual contributions to the NMR chemical shift depend differently on the density functional than those calculated for compound 1. For C-2', the NR part remains constant, the SR contribution varies slightly, and that of SO varies strongly (Figure 8). For N-1, the most constant contribution is surprisingly that of SO, while the NR and SR contributions vary comparably (Figure 9).

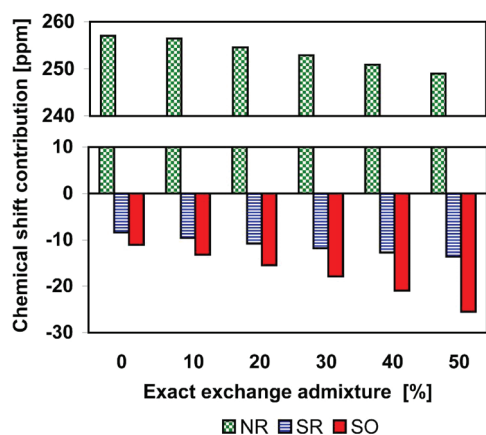


Figure 7. NR, SR, and SO contributions to the NMR chemical shift for N-1 in compound **1** calculated by using a PBE0 functional with various amounts of exact exchange.

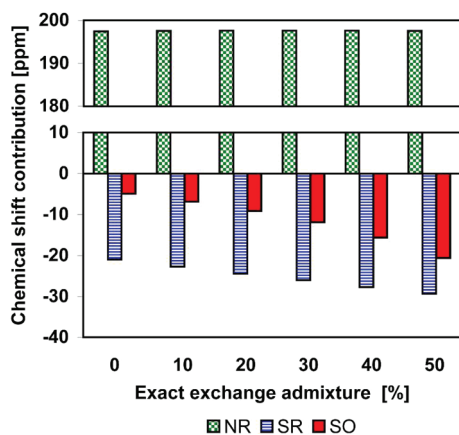


Figure 8. NR, SR, and SO contributions to the NMR chemical shift for C-2' in compound **2** calculated by using a PBE0 functional with various amounts of exact exchange.

3.2.5. Interpretation of the SO Contribution in **1 vs **2**.** In compound **1**, the SO contribution exceeds the SR contribution by a factor of 3 for C-2' and by 1.5 for N-1 (Table 4). In compound **2**, on the contrary, the SR contribution exceeds the SO contribution by a factor of 2 for C-2' and by 2 orders of magnitude for N-1 (-21.1 vs -0.2 ppm, cf. Table 5)! Especially striking is the vanishingly small Δ_{SO} for the Au^{3+} complex. The natural localized MO analysis of the Pt/Au–N bond (14% on the metal, 79% on the nitrogen and equivalent $s-p$ hybridization for the nitrogen for both complexes) does not provide any basis for understanding this phenomenon. However, a significant difference in bonding is reflected in the canonical Kohn–Sham MOs.

Already based on qualitative MO considerations, the bonding in complex **1** should be different from that in complex **2**. The d orbitals should be pushed down for Au^{3+} as compared to Pt^{2+} for two reasons: the higher nuclear charge sensed by the electrons and the presence of the second electronegative chlorine atom.¹⁰⁹ Consequently, the metal d orbitals should become closer in energy to the ligand orbitals, whose weights in the resulting frontier MOs should increase. Indeed, the metal–chlorine π -antibonding highest occupied MO (HOMO) is centered on the metal much more in the case of **1** than in the case of complex **2**.

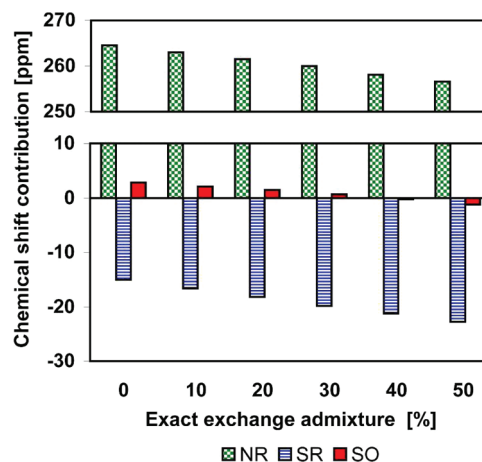


Figure 9. NR, SR, and SO contributions to the NMR chemical shift for N-1 in compound **2** calculated by using a PBE0 functional with various amounts of exact exchange.

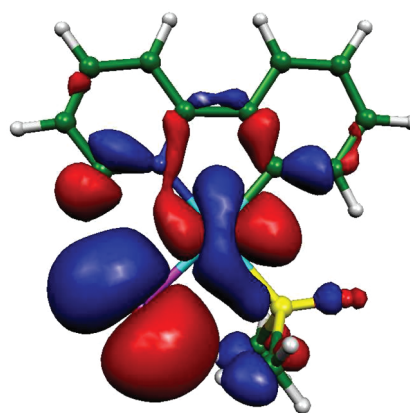


Figure 10. The highest MO relevant for the Δ_{SO} contribution to atom N-1 (HOMO-3) in complex **1**.

The contribution of the HOMO to the net Mulliken population of metal d orbitals is as much as 0.39 for Pt and only 0.12 for Au.

For an orbital to be relevant to the Δ_{SO} contribution, it must contain at least some s character from the light atom in question as well as a contribution from the heavy atom. For reasons of symmetry, the highest MO relevant here is the HOMO-3 (MO lying three levels below the HOMO) of complex **1** (Figure 10). Its counterpart in complex **2** is HOMO-2 (Figure 11). The pattern observed here is the same as for the HOMOs: much more metal character in the Pt^{2+} than in the Au^{3+} complex. In particular, the HOMO-3 of complex **1** contains 28% of the net Pt d and p orbitals and 3% of the net N-1 sp hybrid orbital, whereas the HOMO-2 of complex **2** contains only 12% of the net Au d and p orbitals and 5% of the net N-1 sp hybrid orbital (compare Figures 10 and 11). In both cases, the remainder of the electron density is found mainly in the lone pairs of the one (two) chlorine atom(s). We conclude by stating that the Δ_{SO} for N-1 is much smaller in the Au^{3+} complex due to the higher ionicity of the bonding, since the much smaller metal d-orbital character of the relevant MO induces a smaller “spin polarization” within the nitrogen s orbitals. In contrast to the Δ_{SO} for N-1, the Δ_{SO} for C-2' is *not* quenched in the case of Au^{3+} , presumably because the relevant MO describing the $\text{Au}\cdots\text{C-2'}$ bonding

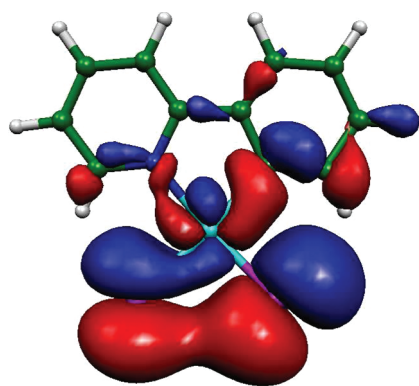


Figure 11. The highest MO relevant for the Δ_{SO} contribution to atom N-1 (HOMO-2) in complex 2.

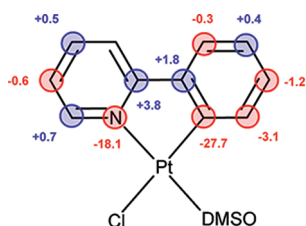


Figure 12. SO contributions to the ^{13}C and ^{15}N NMR chemical shifts for compound 1.

(HOMO-19, 25% of the net Au s and d population, 3% of the net C-2' sp hybrid orbital) lies much lower in energy than HOMO-2. The different behavior of the SO and SR corrections to the N-1 and C-2' shifts in 1 as compared with 2 might also be related to the fact that the corrections for C-2' originate mainly in the in-plane components of the chemical-shift tensor, but the corrections for N-1 are significantly influenced by the out-of-plane δ_{33} component (Table S9, Supporting Information).

The SO contributions to the NMR chemical shifts (Table 4) of the remaining atoms of the aromatic system (Figure 12) are much smaller and, at a first approximation, resemble a “mesomeric effect” spreading from the metal center. The observed trends originate in the propagation of the SO effect via the aromatic system through Fermi-contact (FC) interaction. This propagation resembles the pattern of indirect nuclear spin–spin coupling as has been elegantly demonstrated by Kaupp and co-workers.¹¹⁰

3.2.6. Effects of Geometry. After finding the optimal method for reproducing the experimental ^{13}C NMR chemical shifts using the B3LYP geometry, we then adopted an inverse approach: The optimal PBE0(30)/TZP/ZORA-SO/COSMO method for compound 1 was applied to various optimized geometries (BLYP, B3LYP, BHandHLYP, MP2) and to the X-ray geometry with B3LYP optimized hydrogens. The ^{13}C NMR chemical shifts for the MP2 and X-ray geometries of compound 1 are plotted and compared to experimental NMR data in Figure 13; an analogous comparison is presented for the DFT geometries in Figure S4, Supporting Information.

As our main reference, we again consider the trend in the experimental chemical shifts qualitatively (with the same ordering of carbons according to decreasing chemical shift from theory and experiment) as well as quantitatively (close slopes of the corresponding lines). The PBE0(30)/TZP/ZORA-SO/COSMO approach provides quantitatively correct trends for all carbon

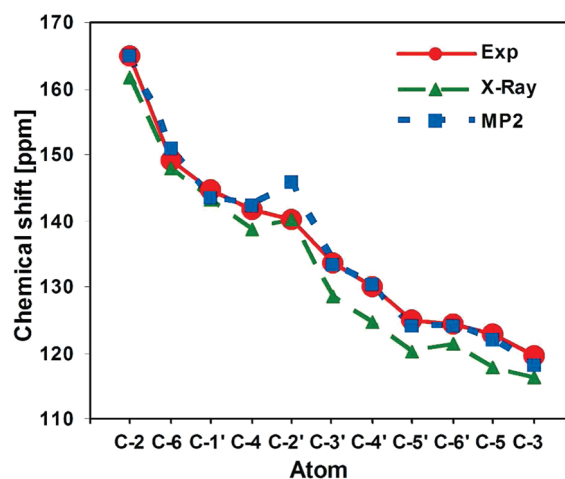


Figure 13. Experimental ^{13}C NMR chemical shifts for compound 1 and those calculated using PBE0(30) for X-ray (with B3LYP optimized protons) and MP2 geometries.

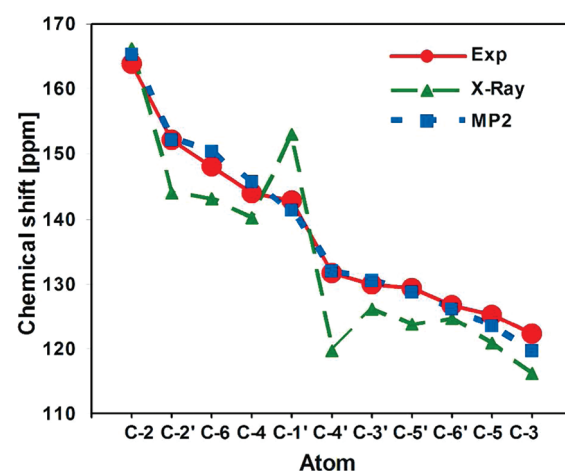


Figure 14. Experimental ^{13}C NMR chemical shifts for compound 2 and those calculated using PBE0(40) for X-ray (with B3LYP optimized hydrogens) and MP2 geometries.

shifts with any of the BLYP, B3LYP, and BHandHLYP structures. The best absolute shifts were found for the B3LYP structure (cf. Figure S4, Supporting Information).

The X-ray structure of 1 provides relatively good agreement with experiment, consistently with a relatively good structural factor. Nevertheless it is less suitable than any of the optimized structures. It overestimates the absolute shieldings for C-3' through C-3, presumably because shorter bond lengths are found within the aromatic rings for the crystal structure than for the optimized geometries. More importantly, the X-ray structure, as well as the MP2 structure, incorrectly describes the qualitative trend between the C-4 and C-2' shifts. Apparently, there is a structural difference between the DFT-optimized structures on the one side and the X-ray and MP2 structures on the other. Indeed, the C-2'-Pt distance is 0.02 Å shorter for the X-ray structure and as much as 0.04 Å shorter for the MP2 structure than what is found by the B3LYP method.

The best absolute shifts were found for the B3LYP structure and, disregarding C-2', also for the MP2 structure. Good quantitative trends were also obtained for the BLYP and BH and HLYP

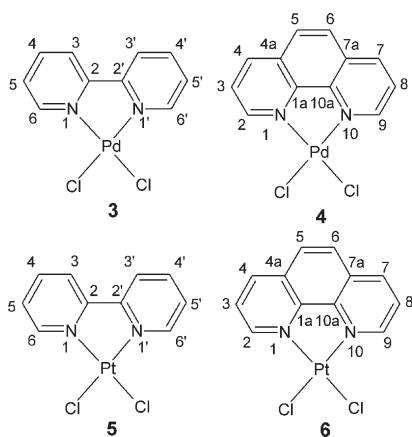
structures. Since the best agreement with NMR experiment was obtained for the B3LYP geometry, the use of this structure for optimizing the method to calculate NMR chemical shifts proved reasonable, and our approach can be considered to be “self-consistent”. A summary of all of the effects gradually included in the calculations is shown schematically for the B3LYP geometry in Figure S5, Supporting Information.

An analogous study of the influence of the geometry (for the DFT employing only the two hybrid functionals) was performed for compound **2**. The PBE0(40)/TZP/ZORA-SO/COSMO results are shown in Figure 14 and Figure S6, Supporting Information. Obviously, the X-ray structure with B3LYP optimized hydrogens (see Section 2.1) shows a dramatic discrepancy with the experimental results for carbons C-1' and C-4'. The NMR chemical shift of C-1' is too large, whereas that of C-4' is too small. The absolute nuclear shielding is thus underestimated for C-1' and overestimated for C-4'. This is consistent with the type of distortion of the benzene ring present in the X-ray structure as compared to the optimized geometry. The bonds from C-1' are elongated,

while those from C-4' are shortened. This could also possibly be due to the presence of two slightly shifted positions of the ligand in the crystal, which are, however, averaged to a single set of atomic positions. Since the *R*-factor for this X-ray structure is relatively large and since all optimized structures provide qualitatively correct trends, we conclude that the X-ray structure of **2** is useless for the calculation of the nuclear magnetic shielding in solution.

The best absolute agreement with experiment was obtained for the less shielded atoms C-2 through C-4 with the BHandHLYP structure (Figure S6, Supporting Information) and for the more shielded atoms C-1' through C-3 with the B3LYP and MP2 geometries. This dependence of the best method for the optimization of structure on the region of shifts in question may be connected to the influence of the solvent: For the most shielded atoms C-5' through C-3, the influence of the COSMO is relatively large. Therefore, an explicit treatment of the solvent, such as a combined molecular dynamics–DFT approach^{111–115} (which is beyond the scope of the current contribution) might unify the best structural method for both regions of the shift.

Chart 2. Structures and Atom Numbering for Compounds 3–6^a



^a[Pd(bpy)Cl₂], **3**; [Pd(phen)Cl₂], **4**; [Pt(bpy)Cl₂], **5**; and [Pt(phen)Cl₂], **6**.

3.3. Applications to Compounds 3–6. In addition to **1** and **2**, we tested the “optimized” computational approach on some other compounds shown in Chart 2: **3**, [Pd(bpy)Cl₂];^{116–118} **4**, [Pd(phen)Cl₂];¹¹⁶ **5**, [Pt(bpy)Cl₂];^{116,118} and **6**, [Pt(phen)Cl₂].¹¹⁶ In the case of Pt and Pd (having the same +2 oxidation state), 30% of the exact exchange admixture for PBE0 functional was applied, whereas for the Au species (+3 oxidation state) we used 40% of the exact exchange admixture. The resulting chemical shifts calculated for compounds **3–6** are summarized in Tables 6 and 7.

It should be noted that while the data for nitrogen and carbon atoms bonded directly to the metal center converge closely to the experimental values when relativistic corrections are applied, the data for carbon atoms more distant from the metal center improve only slightly. In contrast, the NR COSMO data for hydrogen atoms actually deteriorate slightly as compared to the experimental values when relativistic corrections are applied. This deterioration originates, in our opinion, in the insufficient description of the solvation by the implicit solvent model applied here.¹¹⁹ Applications of the explicit solvent model combining classical molecular

Table 6. NMR Chemical Shifts for Compounds 3 and 5 Calculated by ADF using PBE0(30)/TZP/ZORA-SO/COSMO with SR (Δ_{SR}) and SO (Δ_{SO}) Contributions Shown Separately

atom	[Pd(bpy)Cl ₂], 3						[Pt(bpy)Cl ₂], 5					
	δ_{NR}^a	Δ_{SR}	$\delta_{\text{NR}}^b + \sigma_{\text{SR}}$	Δ_{SO}	$\delta_{\text{NR}}^b + \Delta_{\text{SR}} + \Delta_{\text{SO}}$	expt ^e	δ_{NR}^c	Δ_{SR}	$\delta_{\text{NR}}^d + \Delta_{\text{SR}}$	Δ_{SO}	$\delta_{\text{NR}}^d + \Delta_{\text{SR}} + \Delta_{\text{SO}}$	expt ^e
N-1	244.8	−5.0	239.8	−20.5	219.3	217.2	243.5	−7.1	236.4	−31.9	204.5	202.1
C-2	156.1	−1.1	155.0	+1.2	156.2	156.8	156.2	−3.7	152.5	+4.7	157.2	157.3
C-3	123.3	−0.3	123.0	−0.1	122.9	124.3	122.9	−1.3	121.6	+1.0	122.6	124.7
C-4	142.0	−0.1	141.9	+0.1	142.0	141.6	141.9	−1.2	140.7	+1.4	142.1	141.0
C-5	127.4	−0.2	127.2	−0.5	126.7	127.7	127.2	−0.6	126.6	+0.1	126.7	128.2
C-6	151.2	−0.5	150.7	−0.6	150.1	150.1	151.2	−2.3	148.9	+0.5	149.4	148.9
RMSD	0.8		1.1		0.8		1.5		2.6		1.3	
H-3	8.14	−0.08	8.06	0.00	8.06	8.58	8.10	+0.08	8.18	−0.20	7.98	8.58
H-4	8.06	+0.07	8.13	−0.14	7.99	8.35	8.08	−0.03	8.05	+0.03	8.08	8.41
H-5	7.47	0.00	7.47	−0.09	7.38	7.80	7.58	+0.02	7.60	−0.09	7.51	7.84
H-6	8.90	−0.04	8.86	−0.06	8.78	9.12	9.15	−0.06	9.09	+0.16	9.25	9.50
RMSD	0.33		0.35		0.42		0.36		0.36		0.40	

^aADF/PBE0(30)/TZP/COSMO//G03/B3LYP/ECP28MDF/cc-pwCVTZ-PP+6-31G**/PCM. ^bADF/PBE0(30)/TZP/ZORA-SO/COSMO//G03/B3LYP/ECP28MDF/cc-pwCVTZ-PP+6-31G**/PCM. ^cADF/PBE0(30)/TZP/COSMO//G03/B3LYP/ECP60MDF/cc-pwCVTZ-PP+6-31G**/PCM. ^dADF/PBE0(30)/TZP/ZORA-SO/COSMO//G03/B3LYP/ECP60MDF/cc-pwCVTZ-PP+6-31G**+/PCM. ^e¹H and ¹⁵N NMR data: Pazderski, L.; Szyk, E.; Sitkowski, J.; Kamiński, B.; Kozerski, L.; Tousek, J.; Marek, R. *Magn. Reson. Chem.* **2006**, *44*, 163. ¹³C NMR data: this work.

Table 7. NMR Chemical Shifts for Compounds **4** and **6** Calculated by ADF using PBE(30)/TZP/ZORA-SO/COSMO with SR (Δ_{SR}) and SO (Δ_{SO}) Contributions Shown Separately

atom	[Pd(phen)Cl ₂], 4						[Pt(phen)Cl ₂], 6					
	δ_{NR}^a	Δ_{SR}	$\delta_{\text{NR}}^b + \Delta_{\text{SR}}$	Δ_{SO}	$\delta_{\text{NR}}^b + \Delta_{\text{SR}} + \Delta_{\text{SO}}$	expt ^e	δ_{NR}^c	Δ_{SR}	$\delta_{\text{NR}}^d + \Delta_{\text{SR}}$	Δ_{SO}	$\delta_{\text{NR}}^d + \Delta_{\text{SR}} + \Delta_{\text{SO}}$	expt ^e
N-1	241.8	-7.0	238.8	-22.0	216.8	215.4	240.2	-8.4	231.8	-30.5	201.3	200.0
C-2	151.3	-0.4	150.9	-0.8	150.1	149.6	151.5	-2.1	149.4	+0.5	149.9	150.5
C-3	125.6	-0.1	125.5	-0.7	124.8	126.4	125.8	-0.7	125.1	-0.2	124.9	126.3
C-4	141.7	-0.1	141.6	+0.2	141.8	139.9	141.4	-1.2	140.2	+1.6	141.8	140.4
C-5	127.7	-0.2	127.5	+0.1	127.6	128.1	127.4	-0.8	126.6	+0.9	127.5	127.8
C-1a	145.7	-1.0	144.7	+1.4	146.1	148.6	146.0	-3.6	142.4	+5.2	147.6	147.2
C-4a	129.9	-0.2	129.7	0.0	129.7	131.0	129.9	-1.3	128.6	+1.1	129.7	130.9
RMSD	1.0		1.3		1.2		1.6		2.8		1.1	
H-2	9.16	-0.04	9.12	-0.10	9.02	9.34	9.40	-0.06	9.34	+0.01	9.35	9.69
H-3	7.90	-0.01	7.91	-0.08	7.83	8.13	7.95	+0.04	7.99	-0.10	7.89	8.17
H-4	8.70	-0.01	8.69	0.00	8.69	8.97	8.71	-0.04	8.67	-0.02	8.69	9.04
H-5	8.14	0.00	8.14	-0.05	8.09	8.28	8.15	-0.01	8.16	-0.07	8.09	8.29
RMSD	0.21		0.22		0.28		0.26		0.28		0.30	

^a ADF/PBE0(30)/TZP/COSMO//G03/B3LYP/ECP28MDF/cc-pwCVTZ-PP+6-31G**/PCM. ^b ADF/PBE0(30)/TZP/ZORA-SO/COSMO//G03/B3LYP/ECP28MDF/cc-pwCVTZ-PP+6-31G**/PCM. ^c ADF/PBE0(30)/TZP/COSMO//G03/B3LYP/ECP60MDF/cc-pwCVTZ-PP+6-31G**/PCM. ^d ADF/PBE0(30)/TZP/ZORA-SO/COSMO//G03/B3LYP/ECP60MDF/cc-pwCVTZ-PP+6-31G**/PCM. ^e ¹H and ¹³C NMR data: Pazderski, L.; Szyk, E.; Sitkowski, J.; Kamiński, B.; Kozerski, L.; Tousek, J.; Marek, R. *Magn. Reson. Chem.* **2006**, *44*, 163. ¹³C NMR data: this work.

dynamics and relativistic quantum chemical calculations of NMR chemical shifts^{120,121} will be a topic of our future studies. Neglected rovibrational corrections, another possible source of the deviations in the dominant NR part, are expected to play a less significant role because they will, to a large extent, cancel out in the NMR chemical shift calculations.¹²²

4. CONCLUSIONS

There are always two sources of uncertainty in any quantum chemical study that aims to reproduce and predict spectroscopic parameters: a sufficiently precise structure and a sufficiently precise computational method for the spectroscopy. Our study suggests that in cases where the experimental data map the compound globally and thus provide topological trends for the observed property the comparison of theory and experiment can help to identify the proper structure.

More general conclusions can be drawn from the hierarchy of physical effects that influence the value of the NMR chemical shift. First, the quantitatively most important (tenths of ppm) effect comes from inclusion of the relativistic corrections with their magnitudes, however, strongly dependent on the amount of the exact exchange in the density functional. Second is the proper choice of geometry: using a poorly resolved X-ray structure (as in the case of compound **2**) can result in errors of as much as 15 ppm. The choice of the various theoretical optimizations (BLYP, B3LYP, BHandHLYP, MP2) results in a variation of about 5 ppm, which for the three DFT approaches represents only an incremental shift with all of the nuclear shielding trends preserved. Third in importance is the effect of the continuum solvent model (up to 4 ppm), and last is the decontraction and augmentation of the IGLO-III basis set (up to 2.5 ppm).

Another point of general interest which should be highlighted here arises from the interpretation part. The SO contribution to the nuclear magnetic shielding can be rationalized in terms of the composition of those frontier canonical Kohn–Sham orbitals that have a substantial contribution of s character from the light atom in question as well as a contribution from the heavy atom.

Coming to system-specific conclusions, two X-ray geometries and a series of ab initio structures of six Pd²⁺, Pt²⁺, or Au³⁺ complexes were tested on the basis of theoretical vs experimental ¹³C NMR chemical shifts, starting with the evaluation of the NMR approach for a certain geometry and then checking the optimal NMR approach back against various structures. Our calculations demonstrate that the B3LYP method with small-core ECPs and the cc-pwCVTZ-PP/6-31G** basis set augmented with PCM embedding provides geometries for complexes **1–6** which are suitable for calculating ¹³C chemical shifts very close to the experimental values in terms of both the trends and the absolute values. The same is true for the MP2 method tested on complexes **1** and **2**, with the exception of the chemical shift of C-2' in **1**. Good quantitative trends with offsets of the ¹³C chemical shifts ranging from about +3 ppm to about -3 ppm were also obtained for the BLYP structures (the shifts were overestimated in complex **1**) and BHandHLYP structures (the shifts were underestimated in complexes **1** and **2**). The relatively well-resolved X-ray structure of **1** provides correct qualitative trends, except for the chemical shift of C-2', with individual offsets between 0 and -5 ppm; it can therefore be used for qualitative NMR predictions. In contrast, the insufficiently resolved or conformation-averaged X-ray structure of **2** predicts qualitatively incorrect shifts for C-1' and C-4', related to the distortion of the C-1' - C6' aromatic ring. Thus it cannot be used to calculate quantitative NMR data in solution.

Evaluating the approaches for calculating the chemical shift, the most strongly method-dependent term is the largely negative SO contribution to the chemical shifts of C-2' and N-1. Increasing the amount of the exact exchange from 0 to 50% makes the SO contribution more than 10 ppm more negative for C-2' of **2** and N-1 of **1** and by more than 20 ppm for C-2' of **1**. The total SO contributions for these atoms vary from about -5 ppm to -40 ppm. The situation is completely different for N-1 of **2**, where the SO contribution amounts to 0 ± 3 ppm for any of the exact exchange admixtures. We interpret this finding on the basis of the much more

ionic character of the bonding in the Au³⁺ system as compared to the Pt²⁺ complex.

If the SO contributions are less important and less method-dependent in the Au³⁺ as compared to the Pt²⁺ complex, the opposite is true for the SR term. The SR term, which is described equally well by the relativistic ECPs and the ZORA, is again largest for C-2' and N-1 (about -10 ppm in **1**, between -20 and -30 ppm in **2**). However, the SO and SR terms are also sizable (± 3 to ± 5 ppm) for the atoms separated from the metal by two bonds (C-2, C-1', and C-3'), although they partially compensate for each other. The always positive NR chemical shift for the C-2' and N-1 atoms of both **1** and **2** decreases moderately (0 to 8 ppm) upon going from 0% to 50% of the exact exchange admixture.

Adding the SO, SR, and NR terms together and comparing the total chemical shift with the experimental results, we can tune the optimum amount of the exact exchange in the B3LYP or PBE0 density functional. The latter comes out differently for the Pt complex **1** (30%) than for the Au complex **2** (40–50%). This finding, not uncommon in the literature, enables a “pragmatic choice” of the density functional; it is however unsatisfactory from the theoretical point of view. These complexes can thus be suggested as test examples for local hybrid functionals that tune the exact exchange internally. Their development is motivated by a need for a uniform approach to the various systems.^{123–125}

■ ASSOCIATED CONTENT

S Supporting Information. Interatomic distances around the gold center for compound **2**; experimental ¹³C NMR chemical shifts for compound **1** and those calculated by using BLYP, BP86, and PW91 for B3LYP-optimized geometry and KT1, KT2, and PBE0(30) for B3LYP-optimized geometry; experimental ¹³C NMR chemical shifts for compound **1** and those calculated using PBE0(30) for selected DFT geometries (BLYP, B3LYP, BHandHLYP); experimental ¹³C NMR chemical shifts for compound **1** and those calculated using PBE0(30) incorporating various physical effects; experimental ¹³C NMR chemical shifts for compound **2** and those calculated using PBE0(40) for selected DFT geometries (B3LYP, BHandHLYP); selected interatomic distances obtained by using various optimization approaches; NR NMR chemical shifts for compound **1**; SR NMR chemical shifts for compound **1**; NMR chemical shifts for compound **1** obtained from the ADF program using ZORA-SO and functionals VWN, PW92 BP86, PW91, BLYP, B3LYP, BHandHLYP; NMR chemical shifts for compounds **1** and **2** calculated by ADF/B3LYP/TZP/ZORA-SO/COSMO and ADF/PBE0/TZP/ZORA-SO/COSMO with a varying (0–50%) exact exchange admixture parameter; NR, SR, and SO contributions to the principal components of the NMR chemical shift tensor for compounds **1** and **2**; NR, SR, and SO contributions to the NMR chemical shifts for compounds **1** and **2** calculated by ADF/TZP/ZORA-SO/COSMO with a varying (0–50%) exact exchange admixture; and atomic coordinates for all systems employed in this study. This material is available free of charge via the Internet at <http://pubs.acs.org>.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: marketa@ncbr.chemi.muni.cz; rmarek@chemi.muni.cz.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENT

We thank Stanislav Standara (Brno) and Juha Vaara (Oulu) for providing the Fægri-III basis set for platinum and gold and Michal Straka (Prague), Martin Kaupp (Berlin), and Jaromír Marek (Brno) for several helpful discussions. We also would like to thank reviewers for their valuable comments. The authors gratefully acknowledge the financial support of the Ministry of Education of the Czech Republic (grants MSM0021622413 and LC06030 to MM and RM), the project “CEITEC – Central European Institute of Technology” (CZ.1.05/1.1.00/02.0068) from the European Regional Development Fund, and the Erasmus LLP program (T.P. study stay at Masaryk University). The computational resources were partly provided by Meta-Centrum, Czech Republic (grant MSM6383917201).

■ REFERENCES

- (1) *Metal ions in biological systems*; Sigel, A., Sigel, H., Eds; Marcel Dekker, Inc.: New York, 1996; Vol. 33, pp 269–675.
- (2) Orvig, C.; Abrams, M. J. *Chem. Rev.* **1999**, *99*, 2201–2204.
- (3) Messori, L.; Abbate, F.; Marcon, G.; Orioli, P.; Fontani, M.; Mini, E.; Mazzei, T.; Carotti, S.; O'Connell, T.; Zanello, P. *J. Med. Chem.* **2000**, *43*, 3541–3548.
- (4) Okada, T.; El-Mehasseb, I. M.; Kodaka, M.; Tomohiro, T.; Okamoto, K. I.; Okuno, H. *J. Med. Chem.* **2001**, *44*, 4661–4667.
- (5) El-Mehasseb, I. M.; Kodaka, M.; Okada, T.; Tomohiro, T.; Okamoto, K.; Okuno, H. *J. Inorg. Biochem.* **2001**, *84*, 157–158.
- (6) Edwards, G. L.; Black, D. S. C.; Deacon, G. B.; Wakelin, L. P. G. *Can. J. Chem.* **2005**, *83*, 980–989.
- (7) Edwards, G. L.; Black, D. S. C.; Deacon, G. B.; Wakelin, L. P. G. *Can. J. Chem.* **2005**, *83*, 969–979.
- (8) Parish, R. V.; Howe, B. P.; Wright, J. P.; Mack, J.; Pritchard, R. G.; Buckley, R. G.; Elsome, A. M.; Fricker, S. P. *Inorg. Chem.* **1996**, *35*, 1659–1666.
- (9) Fan, D.; Yang, C. T.; Ranford, J. D.; Lee, P. F.; Vittal, J. J. *Dalton Trans.* **2003**, 2680–2685.
- (10) Henderson, W.; Nicholson, B. K.; Faville, S. J.; Fan, D.; Ranford, J. D. *J. Organomet. Chem.* **2001**, *631*, 41–46.
- (11) Fan, D.; Yang, C. T.; Ranford, J. D.; Vittal, J. J.; Lee, P. F. *Dalton Trans.* **2003**, 3376–3381.
- (12) Lo, V. K. Y.; Kung, K. K. Y.; Wong, M. K.; Che, C. M. *J. Organomet. Chem.* **2009**, *694*, 583–591.
- (13) Beletskaya, I. P.; Kashin, A. N.; Karlstedt, N. B.; Mitin, A. V.; Cheprakov, A. V.; Kazankov, G. M. *J. Organomet. Chem.* **2001**, *622*, 89–96.
- (14) Schnyder, A.; Indolese, A. F.; Studer, M.; Blaser, H. U. *Angew. Chem., Int. Ed.* **2002**, *41*, 3668–3671.
- (15) Del Zotto, A.; Prat, F. I.; Baratta, W.; Zangrando, E.; Rigo, P. *Inorg. Chim. Acta* **2009**, *362*, 97–104.
- (16) Atla, S. B.; Kelkar, A. A.; Puranik, V. G.; Bensch, W.; Chaudhari, R. V. *J. Organomet. Chem.* **2009**, *694*, 683–690.
- (17) *Two-Dimensional NMR spectroscopy*; Croasmun, W. R., Carlson, R. M. K., Eds.; VCH: New York, 1994; pp 619–914.
- (18) Ranconi, L.; Sadler, P. J. *Coord. Chem. Rev.* **2008**, *252*, 2239–2277.
- (19) *Calculation of NMR and EPR Parameters. Theory and Application*, Kaupp, M., Bühl, M., Malkin, V. G., Eds.; Wiley-VCH: Weinheim, Germany, 2004; pp 153–277.
- (20) Vaara, J. *Phys. Chem. Chem. Phys.* **2007**, *9*, 5399–5418.
- (21) Solomon, E. I.; Scott, R. A.; King, R. B. *Computational Inorganic and Bioinorganic Chemistry*; John Wiley & Sons: Singapore, 2009, 55–107.
- (22) Pyykkö, P. Theory of NMR parameters. From Ramsey to relativity, 1953–1983. In *Calculation of NMR and EPR Parameters. Theory and Application*; Kaupp, M., Bühl, M., Malkin, V. G., Eds.; Wiley-VCH: Weinheim, Germany, 2004; pp 7–19.
- (23) Pyykkö, P. *Chem. Rev.* **1988**, *88*, 563–594.
- (24) Malkin, V. G.; Malkina, O. L.; Reviakine, R.; Arbuznikov, A. V.; Kaupp, M.; Schimmelpfennig, B.; Malkin, I.; Repisky, M.; Komorovsky,

- S.; Hrobárik, P.; Malkin, E.; Helgaker, T.; Ruud, K. *MAG-ReSpect*, version 2.3; Universität Würzburg: Würzburg, Germany, 2010.
- (25) With new contributions from: Saue, T.; Visscher, L.; Jensen, H. J. Aa.; Bast, R.; Dyall, K. G.; Ekström, U.; Eliav, E.; Enevoldsen, T.; Fleig, T.; Gomes, A. S. P.; Henriksson, J.; Ilias, M.; Jacob, C. R.; Knecht, S.; Nataraj, H. S.; Norman, P.; Olsen, J.; Pernpointner, M.; Ruud, K.; Schimmelpfennig, B.; Sikkema, J.; Thorvaldsen, A.; Thyssen, J.; Villaume, S.; Yamamoto, S. *DIRAC, a relativistic ab initio electronic structure program*, release DIRAC10; University of Southern Denmark: Odense M, Denmark, 2010.
- (26) Hrobárik, P.; Hrobáriková, V.; Meier, F.; Repiský, M.; Komorovský, S.; Kaupp, M. *J. Phys. Chem. A* **2011**, *115*, 5654–5659.
- (27) Vaara, J.; Manninen, P.; Lantto, P. Perturbational and E. C. P. Calculation of Relativistic Effects in NMR Shielding and Spin–Spin Coupling. In *Calculation of NMR and EPR Parameters. Theory and Application*; Kaupp, M., Bühl, M., Malkin, V. G., Eds; Wiley-VCH: Weinheim, Germany, 2004, pp 209–226.
- (28) Guerra, C. F.; Snijders, J. G.; te Velde, G.; Baerends, E. J. *Theor. Chem. Acc.* **1998**, *99*, 391–403.
- (29) Van Lenthe, E.; Baerends, E. J.; Snijders, J. G. *J. Chem. Phys.* **1993**, *101*, 9783–9792.
- (30) Wolff, S. K.; Ziegler, T.; van Lenthe, E.; Baerends, E. J. *J. Chem. Phys.* **1999**, *110*, 7689–7698.
- (31) van Lenthe, E.; Ehlers, A. E.; Baerends, E. J. *J. Chem. Phys.* **1999**, *110*, 8943–8953.
- (32) van Lenthe, E.; Snijders, J. G.; Baerends, E. J. *J. Chem. Phys.* **1996**, *105*, 6505–6516.
- (33) Krykunov, M.; Ziegler, T.; van Lenthe, E. *J. Phys. Chem. A* **2009**, *113*, 11495–11500.
- (34) Godbert, N.; Pugliese, T.; Aiello, I.; Bellusci, A.; Crispini, A.; Ghedini, M. *Eur. J. Inorg. Chem.* **2007**, *2007*, 5105–5111.
- (35) Kobayashi, M.; Masaoka, S.; Sakai, K. *Acta Crystallogr. E* **2008**, *64*, M1325–U3003.
- (36) Pazderski, L.; Pawlak, T.; Sitkowski, J.; Kozerski, L.; Szlyk, E. *Magn. Reson. Chem.* **2009**, *47*, 932–941.
- (37) Craig, C. A.; Garces, F. O.; Watts, R. J.; Palmans, R.; Frank, A. J. *Coord. Chem. Rev.* **1990**, *97*, 193–208.
- (38) 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, Inc., Gaussian 03*, revision E.01; Gaussian, Inc.: Wallingford, CT, 2004.
- (39) Baerends, E. J.; Autschbach, J.; Bashford, D.; Bérces, A.; Bickelhaupt, F. M.; Bo, C.; Boerrigter, P. M.; Cavallo, L.; Chong, D. P.; Deng, L.; Dickson, R. M.; Ellis, D. E.; van Faassen, M.; Fan, L.; Fischer, T. H.; Fonseca Guerra, C.; Ghysels, A.; Giammona, A.; van Gisbergen, S. J. A.; Götz, A. W.; Groeneveld, J. A.; Gritsenko, O. V.; Grüning, M.; Harris, F. E.; van den Hoek, P.; Jacob, C. R.; Jacobsen, H.; Jensen, L.; van Kessel, G.; Kootstra, F.; Krykunov, M. V.; van Lenthe, E.; McCormack, D. A.; Michalak, A.; Mitoraj, M.; Neugebauer, J.; Nicu, V. P.; Noodleman, L.; Osinga, V. P.; Patchkovskii, S.; Philipsen, P. H. T.; Post, D.; Pye, C. C.; Ravenek, W.; Rodríguez, J. I.; Ros, P.; Schipper, P. R. T.; Schreckenbach, G.; Seth, M.; Snijders, J. G.; Solà, M.; Swart, M.; Swerhone, D.; te Velde, G.; Vernooijs, P.; Versluis, L.; Visscher, L.; Visser, O.; Wang, F.; Wesolowski, T. A.; van Wezenbeek, E. M.; Wiesenekker, G.; Wolff, S. K.; Woo, T. K.; Yakovlev, A. L.; Ziegler, T. *ADF2009.01*; SCM: Amsterdam, The Netherlands.
- (40) Godbert, N.; Pugliese, T.; Aiello, I.; Bellusci, A.; Crispini, A.; Ghedini, M. *Eur. J. Inorg. Chem.* **2007**, 5105–5111.
- (41) Maekawa, M.; Munakata, M.; Kitagawa, S.; Nakamura, M. *Anal. Sci.* **1991**, *7*, 521–522.
- (42) Osborn, R. S.; Rogers, D. *J. Chem. Soc., Dalton Trans.* **1974**, 1002–1004.
- (43) Groom, C. R.; Allen, F. H. *WIREs Comput. Mol. Sci.* **2011**, *1*, 368–376.
- (44) Roothaan, C. C. *J. Rev. Mod. Phys.* **1951**, *23*, 69–89.
- (45) Pople, J. A.; Nesbet, R. K. *J. Chem. Phys.* **1954**, *22*, 571–572.
- (46) Moller, C.; Plesset, M. S. *Phys. Rev.* **1934**, *46*, 618–622.
- (47) Head-Gordon, M.; Pople, J. A.; Frisch, M. J. *Chem. Phys. Lett.* **1988**, *153*, 503–506.
- (48) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785–789.
- (49) Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098–3100.
- (50) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648–5652.
- (51) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 1372–1377.
- (52) Koch, W.; Holthausen, M. C. *A Chemist's Guide to Density Functional Theory*, Wiley-VCH: Weinheim, Germany, 2000; pp 119–136.
- (53) Figgen, D.; Peterson, K. A.; Dolg, M.; Stoll, H. *J. Chem. Phys.* **2009**, *130*, 164108.
- (54) Wood, J. H.; Boring, A. M. *Phys. Rev. B* **1978**, *18*, 2701–2711.
- (55) Andrae, D.; Haeussermann, U.; Dolg, M.; Stoll, H.; Preuss, H. *Theor. Chim. Acta* **1990**, *77*, 123–141.
- (56) Martin, J. M. L.; Sundermann, A. *J. Chem. Phys.* **2001**, *114*, 3408–3420.
- (57) Hay, P. J.; Wadt, W. R. *J. Chem. Phys.* **1985**, *82*, 270–283.
- (58) Wadt, W. R.; Hay, P. J. *J. Chem. Phys.* **1985**, *82*, 284–298.
- (59) Hay, P. J.; Wadt, W. R. *J. Chem. Phys.* **1985**, *82*, 299–310.
- (60) Hariharan, P. C.; Pople, J. A. *Theor. Chim. Acta* **1973**, *28*, 213–222.
- (61) Franchl, M. M.; Petro, W. J.; Hehre, W. J.; Binkley, J. S.; Gordon, M. S.; DeFrees, D. J.; Pople, J. A. *J. Chem. Phys.* **1982**, *77*, 3654–3665.
- (62) Krishnan, R.; Binkley, J. S.; Seeger, R.; Pople, J. A. *J. Chem. Phys.* **1980**, *72*, 650–654.
- (63) Miertuš, S.; Scrocco, E.; Tomasi, J. *Chem. Phys.* **1981**, *55*, 117–129.
- (64) Tomasi, J.; Mennucci, B.; Cammi, R. *Chem. Rev.* **2005**, *105*, 2999–3094.
- (65) Bodenhausen, G.; Ruben, D. *J. Chem. Phys. Lett.* **1980**, *69*, 185–189.
- (66) Bax, A.; Summers, M. F. *J. Am. Chem. Soc.* **1986**, *108*, 2083–2084.
- (67) Willker, W.; Leibfritz, D.; Kerssebaum, R.; Bermel, W. *Magn. Reson. Chem.* **1993**, *31*, 287–292.
- (68) Marek, R.; Lyčka, A. *Curr. Org. Chem.* **2002**, *6*, 35–66.
- (69) Marek, R.; Lyčka, A.; Kolehmainen, E.; Sievänen, E.; Toušek, J. *Curr. Org. Chem.* **2007**, *11*, 1154–1205.
- (70) London, F. *J. Phys. Radium* **1937**, *8*, 397–409.
- (71) Pople, J. A. *Proc. R. Soc London* **1957**, *239*, 541–549.
- (72) Pople, J. A. *Proc. R. Soc London* **1957**, *239*, 550–556.
- (73) Pedrew, J. P. *Phys. Rev. B* **1986**, *33*, 8822–8824.
- (74) Pedrew, J. P. *Phys. Rev. B* **1986**, *34*, 7406–7406.
- (75) Perdew, J. P. In *Electronic Structure of Solids '91*; Ziesche, P., Eschring, H., Eds.; Akademie Verlag: Berlin, Germany, 1991; pp 11–20.
- (76) 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 Press: New York, 1998; pp 81–111.
- (77) Perdew, J. P.; Wang, Y. *Phys. Rev. B* **1992**, *45*, 13244–13249.
- (78) Perdew, J. P.; Ernzerhof, M.; Burke, K. *J. Chem. Phys.* **1996**, *105*, 9982–9985.
- (79) Burke, K.; Ernzerhof, M.; Perdew, J. P. *Chem. Phys. Lett.* **1997**, *265*, 115–120.
- (80) Manninen, P.; Lantto, P.; Vaara, J.; Ruud, K. *J. Chem. Phys.* **2003**, *119*, 2623–2637.
- (81) Manninen, P.; Ruud, K.; Lantto, P.; Vaara, J. *J. Chem. Phys.* **2005**, *122*, 114107.
- (82) Fægri, K., Jr.; Almlöf, J. *J. Comput. Chem.* **1986**, *7*, 396–405.

- (83) Kutzelnigg, W.; Fleischer, U.; Schindler, M. In *NMR Basic Principles and Progress*; Diehl, P., Fluck, E., Günther, H., Kosfeld, R., Seelig, J., Eds.; Springer-Verlag: Berlin, Germany, 1990; Vol. 23, pp 165–262.
- (84) Parr, R. G.; Yang, W. *Density-Functional Theory of Atoms and Molecules*; Oxford University Press: Oxford, 1994, pp 152–157.
- (85) Perdew, J. P.; Wang, Y. *Phys. Rev. B* **1992**, *45*, 13244–13249.
- (86) Vosko, S. H.; Wilk, L.; Nusair, M. *Can. J. Phys.* **1980**, *58*, 1200–1211.
- (87) Keal, T. W.; Tozer, D. J. *J. Chem. Phys.* **2003**, *119*, 3015–3024.
- (88) Mennucci, B.; Tomasi, J. *J. Chem. Phys.* **1997**, *106*, 5151–5158.
- (89) Klamt, A.; Schuurmann, G. *J. Chem. Soc., Perkin Trans. 2* **1993**, 799–805.
- (90) Pye, C. C.; Ziegler, T. *Theor. Chem. Acc.* **1999**, *101*, 396–408.
- (91) van Wullen, C. *Phys. Chem. Chem. Phys.* **2000**, *2*, 2137–2144.
- (92) Standara, S.; Bouzková, K.; Straka, M.; Zacharová, Z.; Hocek, M.; Marek, J.; Marek, R. *Phys. Chem. Chem. Phys.* **2011**, *13*, 15854–15864.
- (93) Flukiger, P.; Portmann, S. *MOLEKEL 4.3*; Swiss Center for Scientific Computing: Manno, Switzerland, 2002.
- (94) Jonas, V.; Thiel, W. *J. Chem. Phys.* **1995**, *102*, 8474–8484.
- (95) Pazderski, L.; Tousek, J.; Sitkowski, J.; Kozerski, L.; Szlyk, E. *Magn. Reson. Chem.* **2007**, *45*, 1045–1058.
- (96) Travnicek, Z.; Popa, I.; Cajan, M.; Herchel, R.; Marek, J. *Polyhedron* **2007**, *26*, 5271–5282.
- (97) Leininger, T.; Nicklass, A.; Kuchle, W.; Stoll, H.; Dolg, M.; Bergner, A. *Chem. Phys. Lett.* **1996**, *255*, 274–280.
- (98) Straka, M.; Kaupp, M. *Chem. Phys.* **2005**, *311*, 45–56.
- (99) Huzinaga, S. *Approximate Atomic Functions*, University of Alberta: Edmonton, Canada, 1971.
- (100) Jensen, F. *Introduction to Computational Chemistry*, 1st ed.; Wiley & Sons Ltd.: Chichester, U.K., 1999; p 155.
- (101) *Extensible Computational Chemistry Environment Basis Set Database*, version 1.2.2; Molecular Science Computing Facility, Pacific Northwest Laboratory: Richland, WA.
- (102) Jensen, F. *J. Chem. Theory Comput.* **2008**, *4*, 719–727.
- (103) Malkin, V. G.; Malkina, O. L.; Salahub, D. R. *Chem. Phys. Lett.* **1996**, *261*, 335–345.
- (104) Kaupp, M.; Malkina, O. L.; Malkin, V. G. *Chem. Phys. Lett.* **1997**, *265*, 55–59.
- (105) Wolff, S. K.; Ziegler, T. *J. Chem. Phys.* **1998**, *109*, 895–905.
- (106) Gilbert, T. M.; Ziegler, T. *J. Phys. Chem. A* **1999**, *103*, 7535–7543.
- (107) Bühl, M. *Chem. Phys. Lett.* **1997**, *267*, 251–257.
- (108) Munzarová, M.; Kaupp, M. *J. Phys. Chem. A* **1999**, *103*, 9966–9983.
- (109) Albright, T. A.; Burdett, J. K.; Whangbo, M.-H. *Orbital Interactions in Chemistry*; Wiley & Sons, Inc.: New York, 1985; pp 82 and 145.
- (110) Kaupp, M.; Malkina, O. L.; Malkin, V. G.; Pyykkö, P. *Chem.—Eur. J.* **1998**, *4*, 118–126.
- (111) Malkin, V. G.; Malkina, O. L.; Steinebrunner, G.; Huber, H. *Chem.—Eur. J.* **1996**, *2*, 452–457.
- (112) Pennanen, T. S.; Vaara, J.; Lantto, P.; Sillanpää, A. J.; Laasonen, K.; Jokisaari, J. *J. Am. Chem. Soc.* **2004**, *126*, 11093–11102.
- (113) Bühl, M.; Grigoleit, S.; Kabrede, H.; Mauschick, F. T. *Chem.—Eur. J.* **2005**, *12*, 477–488.
- (114) Asher, J. R.; Doltsinis, N. L.; Kaupp, M. *J. Am. Chem. Soc.* **2004**, *126*, 9854–9861.
- (115) Přecechtělová, J.; Novák, P.; Munzarová, M. L.; Kaupp, M.; Sklenář, V. *J. Am. Chem. Soc.* **2010**, *132*, 17139–17148.
- (116) Pazderski, L.; Szlyk, E.; Sitkowski, J.; Kamiński, B.; Kozerski, L.; Tousek, J.; Marek, R. *Magn. Reson. Chem.* **2006**, *44*, 163–170.
- (117) Maekwa, M.; Munakata, M.; Kitagawa, S.; Nakamura, M. *Anal. Sci.* **1991**, *7*, 521–522.
- (118) Canty, A. J.; Skeleton, B. W.; Traill, P. R.; White, A. H. *Aust. J. Chem.* **1992**, *45*, 417–422.
- (119) Dračinský, M.; Bouř, P. *J. Chem. Theory Comput.* **2010**, *6*, 288–299.
- (120) Standara, S.; Kulhánek, P.; Marek, R.; Horníček, J.; Bouř, P.; Straka, M. *Theor. Chem. Acc.* **2011**, *129*, 677–684.
- (121) Truflandier, L. A.; Autschbach, J. *J. Am. Chem. Soc.* **2010**, *132*, 3472–3483.
- (122) Vaara, J.; Lounila, J.; Ruud, K.; Helgaker, T. *J. Chem. Phys.* **1998**, *109*, 8388–8396.
- (123) Kaupp, M.; Arbuznikov, A.; Bahmann, H. *Zeitschr. Phys. Chem.* **2010**, *224*, 545–568.
- (124) Arbuznikov, A. V.; Kaupp, M. *J. Chem. Phys.* **2009**, *131*, 084103.
- (125) Arbuznikov, A. V.; Kaupp, M. *Chem. Phys. Lett.* **2007**, *440*, 160–168.

Complete Basis Set Extrapolation and Hybrid Schemes for Geometry Gradients of Noncovalent Complexes

Jiří Černý,^{*,†} Michal Pitoňák,[‡] Kevin E. Riley,[§] and Pavel Hobza^{§,||}

[†]Institute of Biotechnology, Academy of Sciences of the Czech Republic, v. v. i., Videňská 1083, 142 20 Praha 4, Czech Republic

[‡]Department of Physical and Theoretical Chemistry, Faculty of Natural Sciences, Comenius University, Mlynská Dolina, 842 15 Bratislava 4, Slovak Republic

[§]Institute of Organic Chemistry and Biochemistry, Academy of Sciences of the Czech Republic, v. v. i., Flemingovo nám. 2, 166 10 Praha 6, Czech Republic

^{||}Department of Physical Chemistry, Palacký University, 771 46 Olomouc, Czech Republic

ABSTRACT: In this paper, we focus on the performance of popular WFT (MP2, MP2.5, MP3, SCS(MI)-MP2, CCSD(T)) and DFT (M06-2X, TPSS-D) methods in optimizations of geometries of noncovalent complexes. Apart from the straightforward comparison of the accuracy of the resulting geometries with respect to the most accurate, computationally affordable, reference method, we have also attempted to determine the most efficient utilization of the information contained in the gradient of a particular method and basis set. Essentially, we have transferred the ideas successfully used for noncovalent interaction energy calculations to geometry optimizations. We have assessed the performance of the hybrid gradients (for instance, MP2 and CCSD(T) calculated in different basis sets), investigated the possibility of extrapolating gradients calculated with a particular method in a series of systematically built basis sets, and finally compared the extrapolated gradients with the counterpoise(CP)-corrected optimizations, in order to determine which of these approaches is more efficient, in terms of their convergence toward the CBS geometry for the respective calculation cost. Further, we compared the efficiency of the CP-corrected, extrapolated, and hybrid gradients in terms of the rate of convergence with respect to basis set size. We have found that CCSD(T) geometries are most faithfully reproduced by the MP2.5 and MP3 methods, followed by the comparably well performing SCS(MI)-MP2 and MP2 methods, and finally by the worst performing DFT-D and M06 methods. Basis set extrapolation of gradients was shown to improve the results and can be considered as a low-cost alternative to the use of CP-corrected gradients. A hybrid gradient scheme was shown to deliver geometries close to the regular gradient reference. Analogously to a similar hybrid scheme, which nowadays is routinely used for the calculation of interaction energies, such a hybrid gradient scheme can save a huge amount of computer time, when high accuracy is desired.

1. INTRODUCTION

The crucial role of noncovalent interactions in biology, nanoscience, catalysis, and other scientific disciplines¹ has been clearly stressed in many scientific discussions and studies during past decades. The necessity of using highly sophisticated *ab initio* quantum chemical methods, such as the coupled-cluster (CC) methods at the complete basis set (CBS) limit, for calculation of noncovalent interactions has also been recognized.^{2–4} The failures of otherwise rather successful methods, such as second-order Møller-Plesset perturbation theory (MP2) or the entire family of DFT methods (except for the recently developed functionals by Truhlar and Zhao⁵ and Scuseria et al.⁶) to describe, for instance, the π - π stacking interactions of aromatic systems, has motivated the community of quantum chemists to develop novel, more efficient (both in terms of accuracy and computational cost) approximate methods. A more detailed discussion of different improved DFT approaches can be found in our review.⁴ Most research efforts have focused solely on interaction energies, and attempts to develop and assess methods that accurately describe noncovalent complex geometries have been largely pushed aside. Except for a few studies on the geometries of hydrogen-bonded (H-bonded) and stacked DNA base pairs^{7–10}

and other complexes,^{11,12} most of the publications in the field have regarded the geometry problem as being less important.

The truth is that the amount of attention paid to the quality of the description of noncovalent complex geometries, compared to interaction energies, to some extent reflects the relative importance of these issues in real life applications. The sensitivity of the geometry of a noncovalent complex to a particular method (and basis set size) applied for geometry optimization is, with no doubt, lower than that of the “single-point” interaction energy. Presuming that a qualitatively correct method (i.e., capable of describing a particular noncovalent interaction type) is used, convergence toward the (at least qualitatively) correct geometry with respect to the basis set size is rather fast. As the most straightforward example, the Hartree–Fock (HF) method with a very small basis set, such as Pople STO-3G, yields at least semiquantitatively correct geometries for “standard” H-bonded complexes (such as the water dimer) in all cases known to us.¹³ The opposite is clearly true if optimization of, for instance, the benzene dimer is attempted using the same methodology. In this

Received: July 11, 2011

Published: October 17, 2011

case, use of a method capable of properly describing the dispersion interaction (i.e., a correlated method, unlike HF) in combination with a basis set containing diffuse functions is critical.

This raises another question, which concerns the convergence of the geometry with respect to the quality of the method (in terms of being closer to the full configuration interaction result), presuming (as mentioned above) that the investigated method is capable of describing the particular interaction type. As an example, we can ask how accurate the geometry of the benzene dimer π - π stacked complex is when it is optimized with a series of MPX (X being 2, 3, or 4) or CC (CCSD, CCSD(T), ...) methods. This certainly is meaningful, since all of the MPX and CC methods are capable of describing dispersion interactions. Furthermore, gaining the knowledge that, for instance, MP2 (being $\sim N^5$ scaling method with system size “N”) delivers similar accuracy to that of the MP4 or the CCSD(T) methods (both scaling as $\sim N^7$) could save a tremendous amount of computer time.

The basis set dependence of optimized geometries has been studied in several works, of which the most relevant to this study are the papers of Salvador and co-workers^{14,15} and Hobza and co-workers^{7,16–18} and also the papers from the groups of Klopper,¹⁹ Pulay and Janowski,²⁰ Kim,¹¹ and others. The common issue discussed in these papers is the role of Basis Set Superposition Error (BSSE) on the molecular (complex) geometry. As was pointed out for the first time in the paper of Simon et al.¹⁴ and further analyzed in the papers of Salvador and Duran¹⁵ and Paizs et al.,²¹ BSSE affects not only interaction energies but also geometrical parameters. In a similar way to how the BSSE leads to an artificial increase in the stabilization of complexes, it also leads to the shortening of, for example, the length of H bonds¹⁴ and the displacement of stacked aromatic rings^{7,20} etc.

We have several goals in this work: first, to assess the performance of the hybrid (for instance, MP2 and CCSD(T) calculated in different basis sets) gradients; second, to investigate the possibility of extrapolating gradients calculated with a particular method in a series of systematically built basis sets; and finally, to compare the extrapolated gradients with the CP-corrected optimizations, to determine which of these approaches is more efficient, in terms of their convergence toward the CBS geometry for the respective calculation cost.

2. METHODS

2.1. CBS Gradient Optimization. In this work, basis set extrapolations are carried out using the scheme of Helgaker et al.^{22,23} (eq 1), where α is a basis set dependent parameter, E_X and E_{CBS} are the HF and correlation energies for the basis set with largest angular momentum X and for the complete basis set, respectively.

$$E_X^{\text{HF}} = E_{\text{CBS}}^{\text{HF}} + A e^{-\alpha X} \text{ and } E_X^{\text{corr}} = E_{\text{CBS}}^{\text{corr}} + BX^{-3} \quad (1)$$

For the RI-MP2/CBS optimization, the HF and MP2 correlation gradients were, as in the case of energies, extrapolated separately due to the different (exponential vs power) convergence profiles of the procedure. The necessary “correlation energy gradient” is obtained as a difference between the total MP2 gradient and the HF gradient. Technically, each element of the HF and correlation Cartesian gradients is extrapolated (using its value instead of energy E in eq 1), and the final CBS gradient is obtained as a sum of the HF/CBS and correlation/CBS gradients. For small complexes, a similar approach was also applied for

higher order correlation methods, directly extrapolating the MP2.5,²⁴ MP3, and CCSD(T) gradients.

2.2. Hybrid Gradient Optimization. The proposed CCSD(T)/CBS gradient is constructed as a sum of the MP2/CBS gradient and a correction gradient, in a way similar to that of the energy calculation. This correction gradient is obtained as a difference between the CCSD(T) and MP2 gradients calculated in a smaller basis set. The 6-31G**(0.25,0.15) and (when applicable) larger cc-pVDZ or cc-pVTZ basis sets were used for the correction gradient calculations. The MP2.5 and MP3 CBS gradients are constructed in a way similar to the one for CCSD(T), employing the (scaled) MP3 correlation gradient correction.

3. COMPUTATIONAL DETAILS

The HF and MP2 CBS calculations were performed with two different consecutive basis sets, cc-pVDZ/cc-pVTZ and cc-pVTZ/cc-pVQZ, using the recommended coefficients for both the energy and gradient element extrapolations.

The optimizations were carried out using our newly developed scripts, which call (depending on the level of theory) the rimp2 module of the TurboMole 5.10 package²⁵ and/or the ACESII/ACESIII²⁶ program for the MP2, MP3 and CCSD(T) energy and gradient calculations. The collected gradients as well as energies are extrapolated, and the relax module of the TurboMole package is called to obtain the new coordinates.

The SCS(MI)-MP2²⁷ optimizations were performed using the recommended parameters of the scs keyword for the ricc2 module of TurboMole.

For a comparison of the performance, we also included the results of DFT calculations, namely, the M06-2X (as implemented in Gaussian 03²⁸) and TPSS-D²⁹ both using the 6-311++G(3df,3pd) basis set. The TPSS-D method is implemented using a combination of the ridft module of TurboMole and our tm_dis program.²⁹

These DFT methods were selected from the wide range of available methods based on our previous results, showing that these methods provide consistently highly accurate interaction energy as well as PES scan results.

All of the optimizations were performed employing the C1 symmetry, even for starting structures with higher symmetry. The same initial geometry was used for all of the optimization methods. The gradients on all atoms of the complex were included, optimizing fully the geometry of monomers within the complex as well as their separation.

Throughout the paper, we are using following notation: D, T, aD, aT, ... for Dunning’s (aug)-cc-pVXZ basis sets; DT and TQ stand for the (X→Y) Helgaker type of basis set extrapolation. “025015” stands for the above-mentioned 6-31G**(0.25, 0.15) modified Pople’s basis set, where the exponent of the most diffuse d functions on heavy (C, N, O, ...) atoms is changed from 0.8 to 0.25, and the exponent of the p function of hydrogens is changed from 1.1 to 0.15. Further, for a particular method “X”, the correction term Δ is defined as $\Delta = X - \text{MP2}$.

The computational cost of optimizations is reduced significantly, especially by employing the hybrid scheme. For example, one cycle of methanol dimer optimization at the CP-CCSD(T)/cc-pVTZ level takes about 60 h, while the hybrid MP2(TQ)+ Δ (CCSD(T)/025015) offers better final geometry with about 2 h for one optimization cycle.

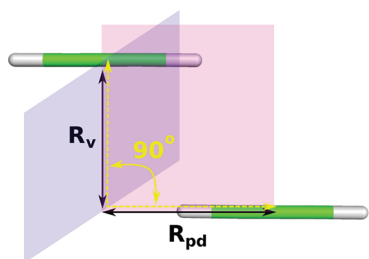


Figure 1. Acetylene dimer, parallel-displaced conformation: structure and the analyzed geometrical parameters.

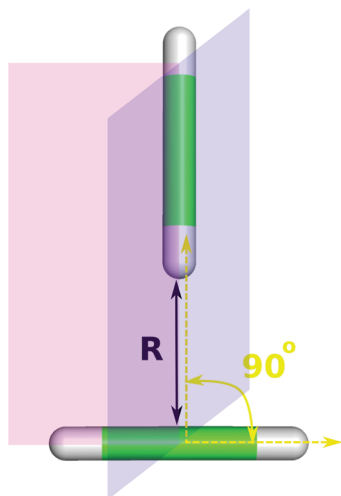


Figure 2. Acetylene dimer, T-shaped conformation: structure and the analyzed geometrical parameters.

4. RESULTS AND DISCUSSION

4.1. Acetylene Dimer. The acetylene dimer was chosen as the model system for π - π stacked, dispersion dominated noncovalent complexes. In analogy with the benzene dimer,^{20,30} two characteristic minima were studied, i.e., the “parallel-displaced”, PD (see Figure 1), and the “T-shaped” structure, T (see Figure 2). Due to the spatial symmetry restrictions applied, the two intermolecular distances, vertical displacement, R_v , and parallel displacement, R_{pd} , define unambiguously the PD structure of the acetylene dimer. In the case of the T structure, it is only the vertical displacement, R , which defines the structure, for the reasons mentioned above. We will repeat here that the monomer geometries were in all cases optimized as well.

The proof of concept for the hybrid MP2/CCSD(T) gradient is demonstrated in Table 1.

Comparing the first two rows of the table, it is clear that MP2 delivers a tighter conformation of the complex compared to CCSD(T), i.e., monomers are vertically shifted by about 0.03 Å further apart at the CCSD(T) level, due to the well-known overestimation of the π - π stacked interaction at the MP2 level. The vertical displacement decreases upon the extension of basis set size (i.e., cc-pVDZ (“D”) to cc-pVTZ (“T”)) at both levels, by 0.26 Å at the MP2 and 0.37 Å at the CCSD(T) level. To demonstrate the success of using the “hybrid gradient”, i.e., mixing of the MP2 gradient calculated in a larger basis set, T, and the CCSD(T) correction to the MP2 gradient (i.e., Δ CCSD(T)) in a smaller basis set, D, we can compare the vertical

Table 1. Acetylene Dimer–PD Conformation^a

MP2/X	Δ CCSD(T)/X	R_v [Å]	R_{pd} [Å]	RMSD [Å]
D		2.8577	3.1947	0.0234
D	D	2.8900	3.2020	0.0311
T		2.8317	3.1492	0.0191
T	D ^b	2.8710	3.1684	0.0095
T	T	2.8534	3.1791	–

^a Comparison of the geometrical parameters obtained on the MP2 level (i.e., no Δ CCSD(T)/X), CCSD(T) level (i.e., MP2/X + Δ CCSD(T)/X), and hybrid MP2/CCSD(T) level (i.e., MP2/X + Δ CCSD(T)/(X-1)). “D” and “T” stand for Dunning’s cc-pVDZ and cc-pVTZ basis sets. RMSD stands for the root mean square deviation of the optimized geometry with regard to the reference geometry (i.e., RMSD of “–”).
^b The 025015 basis set for the Δ CCSD(T)/X calculation performs slightly worse than the D basis set: R_v , 2.8626 [Å]; R_{pd} , 3.1908 [Å]; RMSD, 0.0164 [Å]

displacement of 2.8710 Å, as obtained by the hybrid scheme, with the “pure” MP2 and CCSD(T) values. This value is clearly smaller than that of CCSD(T)/D, in agreement with the trends mentioned above, but also clearly larger than that of the MP2/T value, in accordance with the effect of the higher-order correlation on the gradient. At the same time, this value is in between the CCSD(T)/D and CCSD(T)/T ones, which is a clear improvement, taking into consideration that the computational demands of the hybrid scheme are practically equal to that of pure CCSD(T)/D optimization.

Table 2 summarizes the optimized geometry parameters and the CCSD(T)/CBS interaction energies (calculated as MP2/(aTaQ) + Δ CCSD(T)/cc-pVTZ), calculated for the most accurate geometries obtained by a particular method.

Since obtaining the CCSD(T)/CBS interaction energies for all geometries would be too time-consuming, the discussion on the accuracy of the obtained geometries with respect to the quality of the basis set will be limited only to the MP2 method and can be found in the last paragraph of this section. Let us start first with the T-shaped acetylene dimer. When the MP2/(DT)+ Δ CCSD(T)/025015 (or MP2/(DT)+ Δ CCSD(T)/D) is compared with MP2/(DT) values of the vertical displacement (R), a similar increase of R by about 0.04 (0.05) Å due to the higher-order correlation is observed. Comparing the value of R calculated on the hybrid MP2/(TQ) level corrected with the one Δ CCSD(T) calculated in 025015, D, and T basis sets with the pure MP2/(TQ) value, almost constant shifts of 0.036, 0.048, or 0.041 Å are obtained. This indicates that, similar to calculation of the interaction energies, MP2 gradients converge more slowly with basis set size than the higher order correlation contributions, as calculated at the CCSD(T) level. It is also interesting to analyze the convergence of the MP2 gradient extrapolation. The MP2/(DT) R value repeats the trend of its decreasing value with basis set size, being significantly lower than the MP2/Q value, i.e., 2.6257 vs 2.6537 Å. At the same time, the MP2/(DT) value is closer to the most accurate MP2 R value, obtained either from the MP2/(TQ), 2.6486 Å, or from the two-point “1/X³” R MP2 values extrapolated from T and Q basis sets, 2.6497 Å, by about 0.01 Å.

In analogy with the interaction energy calculations, we can also inspect the CP-corrected gradient optimization results as the concurrent approach to the extrapolated gradients. Both approaches converge to the same geometries, since in the CBS there is no BSSE. However, complexes of a rather small number of atoms can be optimized in basis sets large enough to approach

Table 2. Acetylene Dimer—Selected Geometrical Parameters Obtained from the Unconstrained Geometry Optimization on Various Theoretical Levels and Basis Sets^a

method	hybrid/basis	T			PD				
		R [Å]	RMSD [Å]	E_{int}	R_v [Å]	R_{pd} [Å]	RMSD [Å]	E_{int}	
CC	MP2(DT ^b)+ Δ^c (025015)	2.6675	0.0126		2.8440	3.1710	0.0274		
	MP2(TQ)+ Δ (025015)	2.6847	0.0029		2.8320	3.1074	0.0058		
	MP2(DT)+ Δ (D)	2.6803	0.0071		2.8531	3.1545	0.0237		
	MP2(TQ)+ Δ (D)	2.6962	0.0045		2.8358	3.0858	0.0122		
	MP2(TQ)+ Δ (T)	2.6899	—	−1.500	2.8357	3.1114	—	−1.382	
	CP D	2.8211	0.0816		<i>d</i>		0.9746		
	CP T	2.7225	0.0147		2.8200	3.2597	0.0556		
	CP MP2(D)+ Δ (025015) ^c	2.8654	0.1047		2.9884	3.3286	0.1347		
	CP MP2(T)+ Δ (025015)	2.7584	0.0375		2.8937	3.2413	0.0690		
MP2.5	D	2.7473	0.0424		2.8782	3.1874	0.0460		
	T	2.6520	0.0238		2.8342	3.1635	0.0222		
	MP2(DT)+ Δ (025015)	2.6563	0.0222		2.8449	3.1419	0.0171		
	MP2(TQ)+ Δ (025015)	2.6760	0.0089		2.8241	3.0845	0.0146		
	MP2(DT)+ Δ (D)	2.6576	0.0219		2.8434	3.1349	0.0147		
	MP2(TQ)+ Δ (D)	2.6759	0.0093		2.8323	3.0647	0.0208		
	MP2(TQ)+ Δ (T)	2.6740	0.0109	−1.484	2.8323	3.0774	0.0163	−1.369	
	CP D	2.7972	0.0660		2.8903	3.2956	0.0909		
	CP T	2.6988	0.0061		2.8532	3.1919	0.0379		
	MP3	D	2.7719	0.0519		2.9068	3.1962	0.0573	
		T	2.6794	0.0143		2.8689	3.1709	0.0353	
MP2(DT)+ Δ (025015)		2.6894	0.0103		2.8726	3.1626	0.0450		
MP2(TQ)+ Δ (025015)		2.7084	0.0082		2.8526	3.1043	0.0226		
MP2(DT)+ Δ (D)		2.6886	0.0114		2.8744	3.1402	0.0279		
MP2(TQ)+ Δ (D)		2.7029	0.0064		2.8557	3.0766	0.0164		
MP2(TQ)+ Δ (T)		2.6996	0.0069	−1.470	2.8527	3.0797	0.0140	−1.364	
CP D		2.8232	0.0764		2.9120	3.3187	0.1055		
CP T		2.7281	0.0160		2.8909	3.2080	0.0559		
SCS(MI)-MP2		Q	2.7097	0.0090		2.8491	3.1539	0.0247	
		CP T	2.6897	0.0045		2.8462	3.1871	0.0351	
	CP Q	2.6662	0.0120	−1.506	2.8557	3.1597	0.0285	−1.351	
MP2	D	2.7257	0.0363		2.8577	3.1947	0.0454		
	T	2.6592	0.0146		2.8317	3.1492	0.0170		
	Q	2.6537	0.0181		2.8084	3.1080	0.0115		
	DT ^f	2.6257	0.0357		2.8078	3.1320	0.0120		
	TQ ^g	2.6486	0.0211	−1.497	2.8168	3.0444	0.0330	−1.373	
	CP D	2.7719	0.0570		2.8636	3.2924	0.0838		
	CP T	2.6864	0.0027		2.8368	3.1848	0.0310		
	CP Q	2.6658	0.0122		2.8105	3.1216	0.0074		
	M06-2X	6-311++G(3df,3pd)	2.6700	0.0238	−1.473	2.8167	2.9372	0.0765	−1.323
DFT-D	TPSS/6-311++G(3df,3pd)	2.8498	0.0801	−1.446	2.9277	3.1255	0.0414	−1.368	

^a E_{int} corresponds to the CCSD(T)/CBS interaction energy calculated at the particular complex geometry. RMSD stands for the root mean square deviation of the optimized geometry with regard to the reference geometry (i.e., RMSD of “—”). ^b D, T, ... stand for Dunning's cc-pVXZ basis sets. DT and TQ stand for the (X→Y) Helgaker type of basis set extrapolation. “025015” stands for the 6-31G**(0.25, 0.15) modified Pople's basis set, where the exponent of the most diffuse d functions on heavy (C, N, O, ...) atoms is changed from 0.8 to 0.25, and the exponent of the p function of hydrogens is changed from 1.1 to 0.15. ^c For a particular method “X”, the Δ term is defined as $\Delta = X - \text{MP2}$. ^d Geometry optimization converged to the “T”-shaped structure. ^e CP corrected MP2 gradient with added CP uncorrected Δ term. ^f Extrapolated R_v , R_v , and R_{pd} values according to the two-point “1/X³” extrapolation for D and T basis sets are 2.6312, 2.8208, and 3.1300 Å. ^g Extrapolated R_v , R_v , and R_{pd} values according to the two-point “1/X³” extrapolation for T and Q basis sets are 2.6497, 2.7914, and 3.0779 Å.

the CBS; thus, it is natural to ask whether the CP-corrected gradients in particular basis sets are systematically more accurate than the extrapolated ones. Comparing, for instance, the MP2/(DT)+ Δ CCSD(T)/D value of R with the CP-corrected

one calculated in the D basis set, i.e., 2.6803 vs 2.8211 Å, it is clear that the first one is much closer to the most accurate, MP2/(TQ)+ Δ CCSD(T)/T, value of 2.6899 Å. It is important to realize that the CP-corrected optimization requires the calculation

of five gradients, one for the dimer (as in the hybrid MP2/CCSD(T) scheme) and also four additional gradients of monomers in supermolecular and monomer AO basis sets. After the two-point “ $1/X^3$ ” extrapolation of the CP-corrected R values from the D and T basis sets is done, the obtained value of 2.6810 Å is close to the reference. But the computational overhead of the CP-corrected CCSD(T) calculation in the T basis set is certainly not worth the effort, at least not in this case. The unsatisfactory accuracy of the CP-corrected gradients is even more obvious at the “pure” MP2 level, where the CP-corrected R value in the Q basis set is even less accurate than the plain MP2/T one. On the other hand, two-point “ $1/X^3$ ” extrapolation of CP-corrected MP2 gradients either from D to T or from T to Q basis sets leads (probably by chance) to about the same values, 2.6504 and 2.6508 Å, which are in a good agreement with the best reference MP2 values.

Concerning the performance of methods other than MP2 or CCSD(T), we can compare the “best” geometries for each method, by either taking the MP2/(DT)+ ΔX /T gradients (ΔX being Δ MP2.5 or Δ MP3) and the two-point extrapolation of the CP-corrected gradients in D and T basis sets (or T and Q, in the case of the SCS(MI)-MP2 method) or just by taking only the plain R value (M06-2X and DFT-D). MP2.5 underestimates the reference MP2/(TQ)+ Δ CCSD(T)/T value of R by only 0.016 Å on the extrapolated-hybrid level, but by 0.033 Å (R being 2.6574 Å) on the extrapolated CP-corrected level. The plain CP-corrected MP2.5 value in the T basis set is in much better agreement, with an error of 0.009 Å, but this is most likely the consequence of error compensation due to a lack of basis set saturation. MP3 clearly delivers more accurate results, the agreement of the extrapolated-hybrid and extrapolated CP-corrected scheme (2.6881 Å) being within -0.01 and 0.002 Å. The minus sign means that the R value is overestimated (larger) compared to the reference. The SCS(MI)-MP2 R values match the reference within 0.02 and 0.03 Å at the CP-corrected Q and extrapolated CP-corrected levels. From the DFT methods, it is clearly M06-2X that is more accurate, the error being 0.02 Å vs 0.16 Å for DFT-D.

In the PD acetylene dimer, we observe very similar trends. Higher-order correlation increases the vertical displacement, affecting the parallel displacement only marginally. Favorable error compensation takes place at the MP2 level, since the optimized R_v converges with the basis set from above. As might be anticipated, the best extrapolated-hybrid MP2.5 results reproduce the CCSD(T) reference more closely than those of MP3, the errors being 0.003/0.03 and 0.017/0.03 Å for R_v/R_{pd} , though the difference is rather small. The performance of the CP-corrected SCS(MI)-MP2 in the Q basis set is also satisfactory, slightly overestimating the reference R_v/R_{pd} values by $-0.02/-0.05$ Å. Both the DFT results deviate significantly from the reference, the errors being somewhat lower for the DFT-D method in this case, i.e., 0.02/0.17 Å for M06-2X and 0.09/0.01 Å for DFT-D.

For the interaction energies, it was found⁷ that the higher-order correlation effects not only converge faster with basis set size but also suffer less from BSSE. In the spirit of this idea, we combine the CP-corrected MP2 gradient with the CP-uncorrected Δ CCSD(T) correction calculated with a smaller basis set, for instance, 025015; i.e., see the CP-MP2/D+ Δ /025015 rows in Table 2. The effect of the higher-order correlation on the CP-corrected MP2 gradient is again the elongation of the R , R_v , and R_{pd} distances. However, in both the T-shaped and the PD

Table 3. Acetylene Dimer T, MP2 Method^a

X	E_{int}/X	$E_{\text{int}}/a\text{TaQ}$	E_{def}/X	E_{def}/TQ	$E_{\text{bind}}/a\text{TaQ}/X$
D	-1.164	-1.748 (-0.094)	0.004	1.598	-1.744 (-0.096)
CP-D	-1.182	-1.718 (-0.064)	0.003	1.585	-1.715 (-0.067)
T	-1.380	-1.669 (-0.015)	0.005	0.065	-1.664 (-0.016)
CP-T	-1.386	-1.665 (-0.011)	0.004	0.063	-1.660 (-0.012)
Q	-1.536	-1.659 (-0.005)	0.005	0.017	-1.654 (-0.006)
CP-Q	-1.537	-1.658 (-0.004)	0.005	0.016	-1.654 (-0.006)
DT	-1.496	-1.627 (0.027)	0.006	0.032	-1.622 (0.026)
TQ	-1.651	-1.654	0.006	0.006	-1.648

^a E_{int}/X : CP-corrected interaction energy in “X” basis set and (CP-)“X” optimized geometry. $E_{\text{int}}/a\text{TaQ}$: CP-corrected interaction energy in aTaQ extrapolation and (CP-)“X” optimized geometry. E_{def}/X and E_{def}/TQ : deformation energy in “X” basis set or TQ extrapolation and (CP-)“X” optimized geometry. $E_{\text{bind}}/a\text{TaQ}/X$: CP-corrected interaction energy in aTaQ extrapolation and (CP-)“X” optimized geometry with corrections for the deformation of monomers calculated in the “X” basis set and the corresponding geometry. Values in parentheses represent the error with respect to the “reference” TQ values in the corresponding column.

acetylene dimers, this elongation is in the same direction as the error of the CP-corrected MP2 gradient, thus making such a hybrid method rather inaccurate. Only when the two-point extrapolation of the CP MP2/ $X + \Delta/025015$ from D to T basis set is carried out are the resulting values for R , R_v , and R_{pd} , i.e., 2.7134, 2.8538, and 3.2045 Å, finally more accurate than the MP2 CP-corrected extrapolated values. In the case of the PD structure, such a hybrid method converges to the PD structure, unlike the CP-corrected CCSD(T)/D gradient, which converged to the T-shaped structure.

As mentioned in the beginning of this section, CP-corrected CCSD(T)/CBS interaction energies (without relaxation energies) were calculated for the “best” geometries for each method. These energies vary only slightly for the T-shape structure, from -1.506 (SCS(MI)-MP2 geometry) to -1.446 (DFT-D geometry) kcal/mol, i.e., by only 0.06 kcal/mol ($\sim 4\%$). For the T-shaped structure, the values varied from -1.382 to -1.323 kcal/mol, i.e., again only by 0.06 kcal/mol ($\sim 4\%$). This clearly demonstrates the insensitivity of the interaction energy to geometry changes, thus making the method used for geometry optimization far less important than the method used for the single-point interaction energy calculation. There might be other molecular properties that are more strongly affected by deviations in geometry, such as spectroscopic parameters, but this is out of the scope of this article.

As noted at the beginning of this section, we could afford to calculate the CCSD(T)/CBS interaction energies only for the “best” geometries for each method. However, to get a deeper insight into the importance of basis set saturation in the gradient and interaction energy calculations, we carried out CP-corrected MP2 calculations of the interaction energy in several basis sets and close to the CBS limit at the geometries optimized at the MP2 level in the same series of basis sets with and without the CP-correction in the gradient. The deformation energy contributions to the binding energy are also calculated in the series of basis sets as well as close to the CBS limit. Table 3 summarizes these result for the T-shaped acetylene dimer.

In the first column, i.e., “ E_{int}/X ”, the CP-corrected MP2 interaction energies calculated in the “X” basis set at the CP-corrected and CP-uncorrected MP2 geometries obtained in the same basis set are shown. The interaction energies range from -1.164 (D)

Table 4. Acetylene Dimer PD, MP2 Method^a

X	E_{int}/X	$E_{\text{int}}/\text{aTaQ}$	E_{def}/X	E_{def}/TQ	$E_{\text{bind}}/\text{aTaQ}/X$
D	-0.928	-1.629 (-0.066)	0.003	1.576	-1.626 (-0.068)
CP-D	-0.951	-1.591 (-0.028)	0.003	1.557	-1.588 (-0.030)
T	-1.243	-1.566 (-0.003)	0.004	0.061	-1.562 (-0.004)
CP-T	-1.246	-1.556 (0.007)	0.004	0.059	-1.552 (0.006)
Q	-1.404	-1.566 (-0.003)	0.005	0.016	-1.561 (-0.003)
CP-Q	-1.405	-1.565 (-0.002)	0.005	0.016	-1.560 (-0.002)
DT	-1.381	-1.542 (0.021)	0.005	0.041	-1.538 (0.020)
TQ	-1.540	-1.563	0.006	0.006	-1.558

^a E_{int}/X : CP-corrected interaction energy in “X” basis set and (CP-)“X” optimized geometry. $E_{\text{int}}/\text{aTaQ}$: CP-corrected interaction energy in aTaQ extrapolation and (CP-)“X” optimized geometry. E_{def}/X and E_{def}/TQ : deformation energy in “X” basis set or TQ extrapolation and (CP-)“X” optimized geometry. $E_{\text{bind}}/\text{aTaQ}/X$: CP-corrected interaction energy in aTaQ extrapolation and (CP-)“X” optimized geometry with corrections for deformation of monomers calculated in “X” basis set and the corresponding geometry. Values in parentheses represent the error with respect to the “reference” TQ values in the corresponding column.

to -1.715 kcal/mol (TQ), i.e., the difference being 0.551 kcal/mol ($\sim 32\%$). For all three basis sets, the interaction energies on the CP-corrected geometries are slightly more attractive, but at most only by 0.018 kcal/mol (in the D basis set). If we calculate MP2/CBS interaction energies at these geometries, i.e., the second column “ $E_{\text{int}}/\text{aTaQ}$ ”, these interaction energies only range from -1.627 (DT) to -1.748 (D) kcal/mol; i.e., the span is reduced to only 0.121 kcal/mol ($\sim 7\%$). If we set the TQ geometry to be the reference one, the span of errors in the interaction energies (in parentheses in the second column) reduces even further below 0.1 kcal/mol (0.094 kcal/mol). This clearly demonstrates the importance of the quality of the underlying basis set used for the geometry and single-point calculations. The third and the fourth columns, i.e., “ E_{def}/X ” and “ E_{def}/TQ ”, show the deformation energies as obtained in the “X” basis set, or in the close-to-CBS limit, “TQ”. For this particular complex, the deformation energies are very small, about 0.006 kcal/mol, and are practically invariant to the basis set size (changes from 0.003 to 0.006 kcal/mol from CP-D to TQ). Much larger values are obtained for the CBS limit of the relaxation energy as calculated from geometries optimized in one particular basis set. Values above 1.5 kcal/mol obtained for D and CP-D structures indicate that the monomer geometries (both in the complex and in the isolated systems) are quite different from those obtained in larger basis sets. Indeed, the distance between the carbon atoms within the monomers changes from 1.230 Å in D to 1.209 Å in the Q basis set. In a similar fashion, the length of the C–H bonds shortens from $1.075/1.077/1.076$ Å in D to $1.061/1.064/1.062$ Å, where the first number corresponds to the C–H bond of the monomer “on top”, according to Figure 2 the most distant from the horizontally placed monomer, the second number refers to the C–H of the same monomer pointing toward the monomer on the bottom, and the third value corresponds to the symmetric C–H bonds of the monomer on the bottom.

Similar effects can be observed for the PD structure, shown in Table 4.

The range of the interaction energies in different basis sets and geometries is now 0.659 kcal/mol ($\sim 42\%$), i.e., from -0.928 (D) to -1.587 kcal/mol (TQ). The CBS interaction energies calculated at different geometries span, again, a much narrower

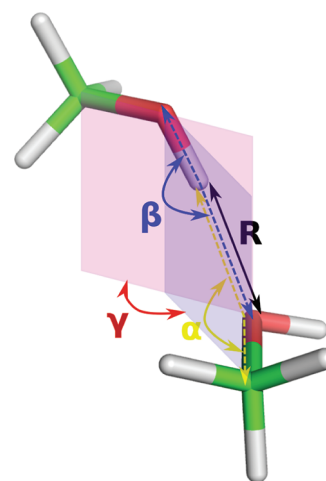


Figure 3. Methanol dimer: structure and the analyzed geometrical parameters.

interval of only 0.088 kcal/mol ($\sim 6\%$), i.e., from -1.538 (DT) to -1.626 (D) kcal/mol. The deformation energies calculated in the same basis set as the gradient are practically identical to those of the T-shaped structure and so is the CBS estimate of the deformation energy for each geometry. For both the T-shaped and the PD structures, the largest error in the MP2/CBS interaction energy (both with and without deformation energy) is obtained for the geometry obtained in the D basis set and is below 0.1 kcal/mol. The errors at the geometries obtained in the T basis set and better are an order of magnitude lower. For such small errors, it is almost meaningless to analyze the deviations for each basis set or extrapolation; nevertheless, the CP-correction to the gradient noticeably increases the accuracy and, at least for the D basis set and the DT extrapolation, does not increase the accuracy compared to the T basis set itself.

4.2. Methanol Dimer. In the H-bonded methanol dimer, three intermolecular geometry parameters will be discussed (see Figure 3).

Parameter “R” represents the length of the H-bond, “ α ” is the angle between the C–O bond of the hydrogen acceptor and the hydrogen involved in bonding, “ β ” is the angle between the O–H bond and hydrogen acceptor oxygen and “ γ ” is the dihedral angle between the C–O and H covalently bound and C–O and H weakly bound defined planes of the two monomers. The results of the all-coordinate optimization are summarized in Table 5.

The analysis of the geometrical parameters is more complicated than in the case of the acetylene dimer because of the large number of intermolecular degrees of freedom. Furthermore, an analysis of such small errors obtained using all of the methods considered here is more or less irrelevant.

Let us start the analysis by comparing the CCSD(T)/CBS interaction energies (calculated as $\text{MP2}/(\text{aTaQ}) + \Delta\text{CCSD(T)}/025015$) at geometries optimized using various methods and basis sets. The best agreement with the reference interaction energy, -5.898 kcal/mol, i.e. the one calculated at the complex geometry obtained using the $\text{MP2}/(\text{TQ}) + \Delta\text{CCSD(T)}/025015$ gradient, is obtained for the $\text{MP2}/(\text{TQ})$ geometry, with an error of only 0.01 kcal/mol. This is in good agreement with the well-known capability of the MP2 method to properly describe (small) H-bonded complexes. The ordering of the rest of the

Table 5. Methanol Dimer—Selected Geometrical Parameters Obtained from the Unconstrained Geometry Optimization on Various Theoretical Levels and Basis Sets^a

method	hybrid/basis	R [Å]	α [deg]	β [deg]	γ [deg]	RMSD [Å]	E_{int}
CC	MP2(DT ^b)+ Δ^c (025015)	1.8888	102.85	163.57	116.93	0.1220	
	MP2(TQ)+ Δ (025015)	1.8876	105.09	167.20	124.79	—	−5.898
	MP2(DT)+ Δ (D)	1.9008	103.84	163.95	118.55	0.0979	
	MP2(TQ)+ Δ (D)	1.9003	106.99	168.10	125.79	0.0362	
	CP D	1.9711	105.02	156.70	76.40	0.6185	
	CP T	1.9233	104.57	160.59	99.08	0.3152	
MP2.5	D	1.8812	102.90	153.06	108.34	0.3825	
	MP2(DT)+ Δ (025015)	1.8890	104.01	164.23	117.66	0.1049	
	MP2(TQ)+ Δ (025015)	1.8887	107.11	167.87	124.21	0.0357	−5.861
	MP2(DT)+ Δ (D)	1.8972	104.84	164.16	119.13	0.1023	
	MP2(TQ)+ Δ (D)	1.8972	108.23	168.12	125.39	0.0608	
	CP D	1.9640	105.42	157.04	77.30	0.6046	
MP3	D	1.8955	103.34	153.48	107.91	0.3745	
	MP2(DT)+ Δ (025015)	1.8979	104.00	164.73	116.51	0.1025	
	MP2(TQ)+ Δ (025015)	1.8977	107.31	169.19	122.70	0.0451	−5.828
	MP2(DT)+ Δ (D)	1.9144	105.65	164.88	119.46	0.0894	
	MP2(TQ)+ Δ (D)	1.9148	109.58	169.26	125.43	0.0808	
	CP D	1.9744	106.01	157.22	78.59	0.5825	
SCS(MI)-MP2	Q	1.8901	107.90	165.97	119.03	0.1111	
	CP T	1.9033	106.93	159.82	95.74	0.3756	
	CP Q	1.9019	108.39	165.14	111.52	0.1725	−5.814
MP2	D	1.8690	102.41	152.62	108.95	0.3888	
	T	1.8754	103.63	160.74	115.80	0.1971	
	Q	1.8773	105.38	165.23	121.06	0.0758	
	DT	1.8798	104.21	163.15	119.23	0.1285	
	TQ	1.8794	106.91	166.91	124.79	0.0506	−5.888
	CP D	1.9550	104.88	156.59	77.60	0.6115	
	CP T	1.9183	104.94	159.05	97.37	0.3569	
	CP Q	1.8958	105.49	163.99	112.60	0.1492	
M06-2X	6-311++G(3df,3pd)	1.9566	103.57	146.46	86.94	0.6136	−5.654
DFT-D	TPSS/6-311++G(3df,3pd)	1.8834	111.72	177.37	126.70	0.1964	−5.943

^a“ E_{int} ” corresponds to the CCSD(T)/CBS interaction energy calculated at the particular complex geometry. RMSD stands for the root mean square deviation of the optimized geometry with regard to the reference geometry (i.e., RMSD of “—”). ^bD, T, ... stand for Dunning’s cc-pVXZ basis sets. DT and TQ stand for the (X→Y) Helgaker type of basis set extrapolation. “025015” stands for the 6-31G** (0.25, 0.15) modified Pople’s basis set, where the exponent of the most diffuse d functions on heavy (C, N, O, ...) atoms is changed from 0.8 to 0.25, and the exponent of the p function of hydrogens is changed from 1.1 to 0.15. ^cFor a particular method “X”, the Δ term is defined as $\Delta = X - \text{MP2}$.

methods, in terms of accuracy, is as follows: MP2.5, DFT-D, MP3, SCS(MI)-MP2, and M06-2X, with errors of −0.037, 0.045, −0.070, −0.084, and −0.244 kcal/mol.

Certain convergence and method-specific patterns in the geometry changes are worth noting here. The convergence of the H-bond length at the CP-uncorrected and CP-corrected MP2 levels is monotonous, but from the opposite direction. CP-uncorrected MP2 H-bond lengths converge from below (1.8690 Å, in D, to 1.8773 Å, in Q basis set), while the CP-corrected bond length converges from above (1.9550 Å, in D, to 1.8958 Å, in Q). Even more distinctive is the change of the dihedral angle “ γ ”, increasing by about 12° at the CP-uncorrected MP2 level, but by about 35° at the CP-corrected level. However, judging by the insensitivity of MP2/CBS interaction energies to these angles in Table 6, the potential energy surface of the dimer is either quite shallow in this direction or the change of the dihedral angle is compensated by the change of other geometrical parameters.

The MP2 interaction energies calculated and optimized with the same basis set vary by 2.39 kcal/mol (~38%), which is comparable to acetylene dimer results. The convergence of the MP2/CBS interaction energies calculated at the (CP-corrected) MP2 geometries in a series of basis sets and basis set extrapolations varies by only 0.19 kcal/mol (3.13%). Inclusion of the deformation energy changes the span of errors only marginally, now being −0.22 kcal/mol (3.84%), smearing out the systematic, though numerically insignificant, higher accuracy of the CP-uncorrected MP2 geometries.

4.3. Benzene···Water Complex. The benzene···water complex is a representative of the O—H··· π interactions and is thus somewhat more challenging to describe than the “standard” H bond in the methanol dimer (Figure 4).

Among all of the possible degrees of freedom in this complex, we chose the distance between the center of mass of the benzene ring and the hydrogen atom involved in the interaction as well as three separate angles. The intermolecular coordinate “ α ” is the

Table 6. Methanol Dimer, MP2 Method^a

X	E_{int}/X	$E_{\text{int}}/a\text{TaQ}$	E_{def}/X	E_{def}/TQ	$E_{\text{bind}}/a\text{TaQ}/X$
D	-3.845	-5.836 (0.069)	0.280	1.157	-5.556 (0.222)
CP-D	-4.098	-5.725 (0.180)	0.138	1.124	-5.587 (0.191)
T	-5.114	-5.909 (-0.004)	0.183	0.225	-5.726 (0.052)
CP-T	-5.098	-5.855 (0.050)	0.141	0.200	-5.713 (0.065)
Q	-5.597	-5.910 (-0.005)	0.140	0.143	-5.770 (0.008)
CP-Q	-5.587	-5.892 (0.013)	0.127	0.134	-5.765 (0.013)
DT	-5.507	-5.900 (0.005)	0.151	0.186	-5.748 (0.030)
TQ	-5.886	-5.905	0.127	0.127	-5.778

^a E_{int}/X : CP-corrected interaction energy in “X” basis set and (CP-)“X” optimized geometry. $E_{\text{int}}/a\text{TaQ}$: CP-corrected interaction energy in aTaQ extrapolation and (CP-)“X” optimized geometry. E_{def}/X and E_{def}/TQ : deformation energy in “X” basis set or TQ extrapolation and (CP-)“X” optimized geometry. $E_{\text{bind}}/a\text{TaQ}/X$: CP-corrected interaction energy in aTaQ extrapolation and (CP-)“X” optimized geometry with corrections for deformation of monomers calculated in “X” basis set and the corresponding geometry. Values in parentheses represent the error with respect to the “reference” TQ values in the corresponding column.

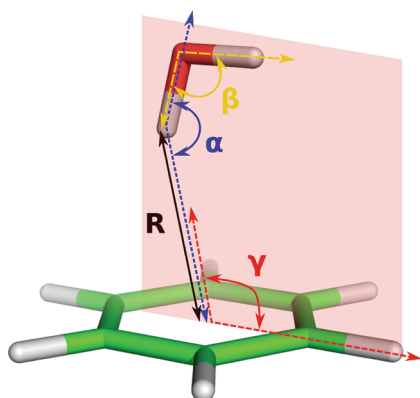


Figure 4. Benzene...water: structure and the analyzed geometrical parameters.

angle between the oxygen, the hydrogen atom involved in the interaction, and the center of mass of the benzene ring, while γ is the angle between the hydrogen atom involved in the interaction, the center of mass of the benzene ring, and the hydrogen atom (of the C–H bond), lying in the intersection of the plane of the water molecule and the benzene ring. Unlike the previous case studies, here we also investigate the intramolecular angle “ β ” formed between the hydrogen atoms and the oxygen in water, which was observed to vary with the method/basis set used in the optimization.

The obtained optimized geometrical parameters along with the CCSD(T)/CBS interaction energies are shown in Table 7.

The basis set dependence of the MP2 gradients and interaction energies is analyzed in Table 8, analogously to the previous test cases.

The CCSD(T)/CBS interaction energies calculated at the geometries obtained with the tested methods again show only a narrow spread of about 0.14 kcal/mol (~4%). Comparing the CCSD(T)/CBS interaction energies calculated at the MP2/(TQ) geometry (Table 7) with the one obtained at the MP2/(TQ) level (Table 8), i.e., -3.35 vs -3.55 kcal/mol, we see that the effect of higher-order correlation is rather small. This can also

be clearly observed in the geometrical parameters, where MP2 delivers essentially the same results as the CCSD(T) method, with the RMSD for DT extrapolation being 0.08 vs 0.09 Å.

One particularly interesting feature of the CP-corrected gradients of the tested methods can be clearly identified. The angle α , describing the bending of the water molecule toward the benzene ring, converges much faster for the CP-corrected gradients compared to the plain ones. For instance, the values of α for the plain MP2 gradients in the D to Q basis sets are 115, 134, and 148°, while the value for CP-corrected gradient in the D basis set is 151°, very close to the CBS limit of about 155°. Concerning the MP2/CBS interaction energies calculated at these geometries (see Table 8), the geometry obtained using the plain gradient in the D basis set appears to be almost converged. Geometries obtained using the CP-corrected gradients, along with all tested basis sets (and extrapolations), deliver geometries for which the MP2/CBS interaction energy differs in hundredths of a kilocalorie per mole.

The SCS(MI)-MP2 method does not perform better than the MP2 method itself, judging from the RMSD values. However, the small elongation of the benzene...water distance caused by the higher-order correlation, e.g., 2.441 Å on the best CCSD(T)/CBS estimate in Table 7 vs 2.428 Å on the MP2/(TQ) level, is reproduced at the spin-component scaled MP2 level, 2.504 vs 2.438 Å for SCS(MI)-MP2/Q vs MP2/Q. The performance of the DFT methods is noticeably worse, but still satisfactory. Unlike M06-2X, the TPSS-D method predicts a significantly larger benzene...water separation, 2.739 Å (+ 0.3 Å from the reference), but both methods clearly underestimate the α angle by as much as approximately 30°.

4.4. Pyrrol Dimer. The largest investigated complex, for which the analytic CCSD(T) geometry gradient is still computationally feasible, is the pyrrol dimer in a T-shape tilted conformation, see Figure 5.

According to the known character of π - π system interaction, the pyrrol dimer clearly has the largest dispersion contribution among the tested complexes. Due to the fact that the π - π dispersion interaction is the most challenging one to describe, the most diverse results for the tested methods are expected. Because of the size of the system, only a limited set of results, in terms of the variety and size of the basis sets, could be obtained, thus offering only a limited space for demonstration of the general trends. The mutual distance between the pyrrol rings is monitored by the parameter “R”, which is the distance between the hydrogen, bound to nitrogen, pointing toward the second pyrrol ring and the center of mass of that second pyrrol ring.

The “ α ” and “ β ” angles then reflect the tendency of the rings to fold against each other, converging to the so-called “sandwich” structure.

To our surprise, the range of the CCSD(T)/CBS interaction energies obtained for “the best” geometries for each test method is again very narrow, less than 0.1 kcal/mol (~1%; Table 9).

Concerning the similarity of the reference MP2/(DT)+ Δ CCSD(T)/025015 geometry with the ones obtained using lower level methods, the best match is obtained using the MP2/(DT)+ Δ MP2.5/025015 hybrid gradient. MP2 and MP3 results are in full agreement with the general trend of MP2 to overestimate the magnitude of the π - π dispersion interaction, thus shortening the intermolecular distances, and MP3 having the opposite effect. SCS(MI)-MP2, known for having a much better description of this type of noncovalent interaction, also delivers results in excellent agreement with CCSD(T), being the

Table 7. Benzene...Water—Selected Geometrical Parameters Obtained from the Unconstrained Geometry Optimization on Various Theoretical Levels and Basis Sets^a

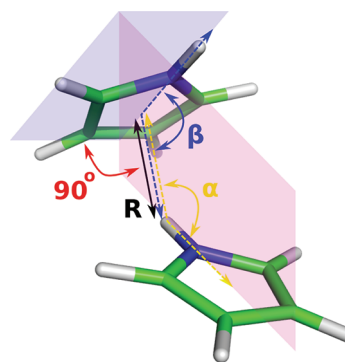
method	hybrid/basis	R [Å]	α [deg]	β [deg]	γ [deg]	RMSD [Å]	E_{int}
CC	MP2(DT ^b)+ Δ^c (025015)	2.4623	138.52	102.97	102.70	0.0927	
	MP2(TQ)+ Δ (025015)	2.4414	152.31	103.67	102.94	—	−3.374
MP2.5	D	2.6960	114.48	100.57	105.95	0.1997	
	MP2(DT)+ Δ (025015)	2.4671	139.47	103.15	103.56	0.0783	
	MP2(TQ)+ Δ (025015)	2.4582	154.12	103.83	103.17	0.0198	−3.346
	MP2(DT)+ Δ (D)	2.4751	140.40	103.19	103.52	0.0706	
	MP2(TQ)+ Δ (D)	2.4688	155.49	103.86	103.15	0.0311	
	CP D	2.6158	151.03	101.26	100.94	0.0703	
MP3	D	2.7363	114.49	100.88	105.72	0.1945	
	MP2(DT)+ Δ (025015)	2.5010	137.98	103.37	103.46	0.0819	
	MP2(TQ)+ Δ (025015)	2.4887	153.33	104.04	103.12	0.0253	−3.336
	MP2(DT)+ Δ (D)	2.5150	140.02	103.44	103.25	0.0681	
	MP2(TQ)+ Δ (D)	2.5110	155.66	104.09	103.03	0.0455	
SCS(MI)-MP2	CP D	2.6539	149.95	101.51	101.20	0.0800	
	Q	2.5042	149.37	104.09	103.28	0.0225	
	CP T	2.5021	153.65	103.76	102.39	0.0301	
MP2	CP Q	2.5152	155.92	104.23	102.93	0.0484	−3.291
	D	2.6540	115.12	100.21	105.96	0.2031	
M06−2X	T	2.4852	134.28	102.16	103.84	0.1102	
	Q	2.4379	148.08	103.10	103.35	0.0284	
	DT	2.4373	140.47	102.91	103.78	0.0777	
	TQ	2.4277	154.88	103.61	103.20	0.0184	−3.346
	CP D	2.5860	151.18	100.94	100.93	0.0622	
	CP T	2.4613	154.49	102.67	101.77	0.0187	
	CP Q	2.4471	155.18	103.26	102.65	0.0174	
	DFT-D	TPSS/6-311++G(3df,3pd)	2.7387	124.75	102.73	104.84	0.1351

^a“ E_{int} ” corresponds to the CCSD(T)/CBS interaction energy calculated at the particular complex geometry. RMSD stands for the root mean square deviation of the optimized geometry with regard to the reference geometry (i.e., RMSD of “−”). ^bD, T, ... stand for Dunning's cc-pVXZ basis sets. DT and TQ stand for the (X→Y) Helgaker type of basis set extrapolation. “025015” stands for the 6-31G** (0.25, 0.15) modified Pople's basis set, where the exponent of the most diffuse d functions on heavy (C, N, O, ...) atoms is changed from 0.8 to 0.25 and the exponent of the p function of hydrogens is changed from 1.1 to 0.15. ^cFor a particular method “X”, the Δ term is defined as $\Delta = X - \text{MP2}$.

Table 8. Benzene...Water, MP2 Method^a

X	E_{int}/X	$E_{\text{int}}/\text{aTaQ}$	E_{def}/X	E_{def}/TQ	$E_{\text{bind}}/\text{aTaQ}/X$
D	−1.509	−3.656 (−0.107)	0.051	1.738	−3.605 (−0.078)
CP-D	−2.203	−3.587 (−0.038)	0.019	1.642	−3.568 (−0.041)
T	−2.735	−3.554 (−0.005)	0.036	0.150	−3.518 (0.009)
CP-T	−2.906	−3.583 (−0.034)	0.019	0.118	−3.564 (−0.037)
Q	−3.250	−3.564 (−0.015)	0.024	0.048	−3.539 (−0.012)
CP-Q	−3.274	−3.562 (−0.013)	0.019	0.042	−3.542 (−0.015)
DT	−3.220	−3.516 (0.033)	0.032	0.086	−3.484 (0.043)
TQ	−3.552	−3.549	0.022	0.022	−3.527

^a E_{int}/X : CP-corrected interaction energy in “X” basis set and (CP-)“X” optimized geometry. $E_{\text{int}}/\text{aTaQ}$: CP-corrected interaction energy in aTaQ extrapolation and (CP-)“X” optimized geometry. E_{def}/X and E_{def}/TQ : deformation energy in “X” basis set or TQ extrapolation and (CP-)“X” optimized geometry. $E_{\text{bind}}/\text{aTaQ}/X$: CP-corrected interaction energy in aTaQ extrapolation and (CP-)“X” optimized geometry with corrections for deformation of monomers calculated in “X” basis set and the corresponding geometry. Values in parentheses represent the error with respect to the “reference” TQ values in the corresponding column.

**Figure 5. Pyrol dimer: Structure and the Analyzed Geometrical Parameters.**

second most accurate one here. The basis set convergence of the MP2 geometries is not monotonous, leading to quite different results for the DT and TQ extrapolations. This is further manifested in noticeably distinct MP2.5 and MP3 results, when used with these extrapolated MP2 gradients. However, the

Table 9. Pyrol Dimer—Selected Geometrical Parameters Obtained from the Unconstrained Geometry Optimization on Various Theoretical Levels and Basis Sets^a

method	hybrid/basis	R [Å]	α [deg]	β [deg]	RMSD [Å]	E_{int}
CC	MP2(DT ^b)+ Δ^c (025015)	2.2554	136.30	95.07	—	−6.664
MP2.5	D	2.2863	133.10	95.85	0.0329	
	MP2(DT)+ Δ (025015)	2.2518	136.00	96.08	0.0131	
	MP2(TQ)+ Δ (025015)	2.2823	132.64	95.93	0.0395	−6.713
	CP D	2.3175	141.60	95.92	0.0985	
MP3	D	2.3254	136.86	95.45	0.0476	
	MP2(DT)+ Δ (025015)	2.2864	140.20	95.66	0.0624	
	MP2(TQ)+ Δ (025015)	2.3080	139.38	95.56	0.0581	−6.635
	CP D	2.3644	145.26	95.54	0.1552	
SCS(MI)-MP2	Q	2.2800	136.32	96.69	0.0242	
	CP T	2.2596	137.04	96.67	0.0279	
	CP Q	2.2836	137.64	96.82	0.0402	−6.656
MP2	D	2.2578	129.18	96.22	0.0842	
	T	2.2350	130.66	96.57	0.0713	
	Q	2.2509	129.32	96.72	0.0858	
	DT	2.2270	130.63	96.55	0.0768	
	TQ	2.2747	125.09	96.40	0.1431	−6.672
	CP D	2.2782	137.53	96.37	0.0421	
	CP T	2.2444	134.30	96.64	0.0237	
CP Q	2.2496	131.86	96.84	0.0516		
M06−2X	6-311++G(3df,3pd)	2.6979	97.64	87.94	0.5564	−6.722
DFT-D	TPSS/6-311++G(3df,3pd)	2.3420	132.84	97.71	0.0391	−6.678

^a E_{int} corresponds to the CCSD(T)/CBS interaction energy calculated at the particular complex geometry. RMSD stands for the root mean square deviation of the optimized geometry with regard to the reference geometry (i.e., RMSD of “−”). ^b D, T, ... stand for Dunning's cc-pVXZ basis sets. DT and TQ stand for the (X→Y) Helgaker type of basis set extrapolation. “025015” stands for the 6-31G**(0.25, 0.15) modified Pople's basis set, where the exponent of the most diffuse d functions on heavy (C, N, O, ...) atoms is changed from 0.8 to 0.25, and the exponent of the p function of hydrogens is changed from 1.1 to 0.15. ^c For a particular method “X”, the Δ term is defined as $\Delta = X - \text{MP2}$.

Table 10. Pyrol Dimer, MP2 Method^a

X	E_{int}/X	$E_{\text{int}}/a\text{TaQ}$	E_{def}/X	E_{def}/TQ	$E_{\text{bind}}/a\text{TaQ}/X$
D	−5.585	−8.174 (−0.160)	0.156	2.331	−8.018 (−0.228)
CP-D	−5.854	−8.025 (−0.011)	0.114	2.237	−7.911 (−0.121)
T	−7.097	−7.988 (0.026)	0.165	0.283	−7.823 (−0.033)
CP-T	−7.144	−7.956 (0.058)	0.149	0.260	−7.807 (−0.017)
Q	−7.591	−7.983 (0.031)	0.177	0.200	−7.807 (−0.017)
CP-Q	−7.589	−7.963 (0.051)	0.161	0.183	−7.803 (−0.013)
DT	−7.754	−7.919 (0.095)	0.175	0.203	−7.744 (0.046)
TQ	−8.035	−8.014	0.224	0.224	−7.790

^a E_{int}/X : CP-corrected interaction energy in “X” basis set and (CP-)“X” optimized geometry. $E_{\text{int}}/a\text{TaQ}$: CP-corrected interaction energy in aTaQ extrapolation and (CP-)“X” optimized geometry. E_{def}/X and E_{def}/TQ : deformation energy in “X” basis set or TQ extrapolation and (CP-)“X” optimized geometry. Values in parentheses represent the error with respect to the “reference” TQ values in the corresponding column.

differences are slightly buffered by the presence of higher-order correlation.

CP-corrected MP2 gradients are superior to the plain ones in all of the tested basis sets; however, MP2.5 and MP3 CP-corrected results in the D basis sets are inferior to those obtained on the computationally cheaper, hybrid MP2/(DT)+ Δ /025015 level.

There are large discrepancies in the quality of the results produced by the two DFT methods tested in this work. While the TPSS+D delivers one of the most accurate results, with a RMSD

value comparable with that of MP2.5/CBS, the M06-2X method converges to the most distinct geometry among all of the tested methods. This method produces the largest displacement of the pyrol rings, with deviations of more than 0.2 Å on average. The α angle is, on the other hand, 30% smaller, resembling the “sandwich” rather than the “T-shaped” structure.

Concerning the detailed basis set convergence of the plain- and CP-corrected MP2 gradients and interaction energies, shown in Table 10, it is clear that for such a large molecular system, basis set saturation is reached much faster.

The MP2/(aTaQ) interaction energy calculated at the plain MP2 gradient geometry in the D basis set is only 0.16 kcal/mol different than the one obtained for the most accurate MP2/(TQ) geometry. When the interaction energies are calculated in the same basis sets that were used for the geometry optimization, CP-corrected gradients are naturally more accurate. Nevertheless, when the CP-corrected T and extrapolated DT gradients are compared, the latter one leads to a more accurate interaction energy, −7.14 vs −7.75 kcal/mol, for a significantly lower computational cost.

5. CONCLUSIONS

We compared the efficiency of the CP-corrected gradients in terms of rate of convergence with respect to basis set size, with extrapolated and hybrid gradients. Essentially, what we have tried to do is transfer the ideas successfully used for noncovariant interaction energy calculations to geometry optimizations.

Furthermore, we have also approached the problem from a pragmatical point of view, seeking to find the level of theory for geometry description that is needed for the subsequent single-point interaction energy calculations (performed at the reliable and accurate CCSD(T) level of theory).

We can generally conclude the following:

1. When the CCSD(T)/CBS (extrapolated MP2 augmented with the higher-order correlation corrections from a CCSD(T)) interaction energy calculated at geometries optimized at various theoretical WFT/DFT levels is used as a measure, only small differences are obtained (less than 5% for the tested complexes).
2. Comparing the RMSD of a complex geometry optimized using a particular WFT/DFT method and basis set with respect to the reference geometry, obtained at the highest computationally feasible level (extrapolated MP2 gradient augmented with the higher-order correlation corrections from CCSD(T)), more pronounced differences are noticeable. CCSD(T) geometries are most faithfully reproduced by the MP2.5 and MP3 methods, followed by the comparably well performing SCS(MI)-MP2 and MP2 methods, finally by the worst performing DFT-D and M06-2X methods.
3. A detailed basis set investigation of the MP2 gradient and interaction energies confirmed the known fact that geometries converge much faster with basis set size than do interaction energies. Geometry optimization using the cc-pVTZ basis set delivers practically converged geometries with respect to the CBS ones.
4. Basis set extrapolation of gradients was shown to improve the results and can be considered as a low-cost alternative to the use of CP-corrected gradients.
5. A hybrid scheme consisting of an MP2 gradient calculated at a larger, "X+1", basis set (or basis set extrapolation) and the post-MP2 (higher-order correlation) gradient in a smaller, "X", basis set was shown to deliver geometries close to those obtained using the plain post-MP2 "X+1" gradient. Analogously to a similar hybrid scheme, which nowadays is routinely used for the calculation of interaction energies, such a hybrid gradient scheme can save a huge amount of computer time, when high accuracy is desired.

AUTHOR INFORMATION

Corresponding Author

*E-mail: jiri.cerny@img.cas.cz.

ACKNOWLEDGMENT

This work was a part of the research project no. Z40550506 of the Institute of Organic Chemistry and Biochemistry, Academy of Sciences of the Czech Republic, and it was supported by Grant Nos. LC512 and MSM6198959216 from the Ministry of Education, Youth and Sports of the Czech Republic. The support of Praemium Academiae, Academy of Sciences of the Czech Republic, awarded to P.H. in 2007 is also acknowledged. This work was also supported by the Korea Science and Engineering Foundation (World Class Univ. program: R32-2008-000-10180-0) and by the Slovak Grant Agency VEGA under the contract No. 1/0520/10 and the Slovak Research and Development Agency under the contract No. APVV-0059-10. J.C. gratefully acknowledges support by the Institutional Research concept No. AV0Z50520701 of the Academy of Sciences of the Czech

Republic, and an important part of the calculations was performed using the "Chinook" supercomputer at the Environmental Molecular Sciences Laboratory of the Pacific Northwest National Laboratory.

REFERENCES

- (1) Lee, E. C.; Kim, D.; Jurečka, P.; Tarakeshwar, P.; Hobza, P.; Kim, K. S. *J. Phys. Chem. A* **2007**, *111*, 3446.
- (2) Sinnokrot, M. O.; Sherrill, C. D. *J. Phys. Chem. A* **2004**, *108*, 10200.
- (3) Šponer, J.; Riley, K. E.; Hobza, P. *Phys. Chem. Chem. Phys.* **2008**, *10*, 2595.
- (4) Riley, K. E.; Pitoňák, M.; Jurečka, P.; Hobza, P. *Chem. Rev.* **2010**, *110*, 5023.
- (5) Zhao, Y.; Truhlar, D. G. *J. Chem. Theory Comput.* **2007**, *3*, 289.
- (6) Janesko, B. G.; Henderson, T. M.; Scuseria, G. E. *J. Chem. Phys.* **2009**, *131*, 034110.
- (7) Dąbkowska, I.; Jurečka, P.; Hobza, P. *J. Comput. Chem.* **2005**, *122*, 204322.
- (8) Riley, K. E.; Pitoňák, M.; Černý, J.; Hobza, P. *J. Chem. Theory Comput.* **2010**, *6*, 66.
- (9) Černý, J.; Hobza, P. *Chem. Commun.* **2010**, *46*, 383–385.
- (10) Černý, J.; Hobza, P. *Phys. Chem. Chem. Phys.* **2005**, *7*, 1624–1626.
- (11) Min, S. K.; Lee, E. C.; Lee, H. M.; Kim, D. Y.; Kim, D.; Kim, K. S. *J. Comput. Chem.* **2008**, *29*, 1208.
- (12) Černý, J.; Tong, X.; Hobza, P.; Müller-Dethlefs, K. *Phys. Chem. Chem. Phys.* **2008**, *10*, 2780.
- (13) Hobza, P.; Zahradník, R. *Weak Intermolecular Interactions In Chemistry and Biology*; Elsevier Scientific Pub. Co.: Amsterdam, 1980.
- (14) Simon, S.; Duran, M.; Dannenberg, J. J. *J. Chem. Phys.* **1996**, *105*, 11024.
- (15) Salvador, P.; Duran, M. *J. Chem. Phys.* **1999**, *111*, 4460.
- (16) Hobza, P.; Havlas, Z. *Theor. Chem. Acc.* **1998**, *99*, 372.
- (17) Hobza, P.; Bludský, O.; Suhai, S. *Phys. Chem. Chem. Phys.* **1999**, *1*, 3073.
- (18) Riley, K. E.; Hobza, P. *J. Chem. Theory Comput.* **2010**, *6*, 66.
- (19) Boese, A. D.; Jansen, G.; Torheyden, M.; Höfener, S.; Klopper, W. *Phys. Chem. Chem. Phys.* **2011**, *13*, 1230.
- (20) Janowski, T.; Pulay, P. *Chem. Phys. Lett.* **2007**, *447*, 27.
- (21) Paizs, B.; Salvador, P.; Császár, A. G.; Duran, M.; Suhai, S. *J. Comput. Chem.* **2001**, *22*, 196.
- (22) Halkier, A.; Helgaker, T.; Jørgensen, P.; Klopper, W.; Koch, H.; Olsen, J.; Wilson, A. K. *Chem. Phys. Lett.* **1998**, *286*, 243.
- (23) Halkier, A.; Helgaker, T.; Jørgensen, P.; Klopper, W.; Olsen, J. *Chem. Phys. Lett.* **1999**, *302*, 437.
- (24) Pitoňák, M.; Neogrády, P.; Černý, J.; Grimme, S.; Hobza, P. *Chem. Phys. Chem.* **2009**, *10*, 282.
- (25) Ahlrichs, R.; Bär, M.; Häser, M.; Horn, H.; Kölmel, C. *Chem. Phys. Lett.* **1989**, *162*, 165–169.
- (26) Lotrich, V.; Flocke, N.; Ponton, M.; Yau, A. D.; Perera, A.; Deumens, E.; Bartlett, R. J. *J. Chem. Phys.* **2008**, *128*, 194104.
- (27) DiStasio, R. A.; Head-Gordon, M. *Mol. Phys.* **2007**, *105*, 1073.
- (28) Frisch, M. J. et al. *Gaussian 03*; Gaussian Inc.: Wallingford CT, 2003.
- (29) Jurečka, P.; Černý, J.; Hobza, P.; Salahub, D. R. *J. Comput. Chem.* **2007**, *28*, 555–569.
- (30) Pitoňák, M.; Neogrády, P.; Řezáč, J.; Jurečka, P.; Urban, M.; Hobza, P. *J. Chem. Theory Comput.* **2008**, *4*, 1829.

Evaluation of the Nonlinear Optical Properties for Annulenes with Hückel and Möbius Topologies

Miquel Torrent-Sucarrat,^{*,†} Josep M. Anglada,[†] and Josep M. Luis[‡]

[†]Departament de Química Biològica i Modelització Molecular, Institut de Química Avançada de Catalunya (IQAC–CSIC), c/Jordi Girona 18, E-08034 Barcelona, Spain

[‡]Institut de Química Computacional and Departament de Química, Universitat de Girona, E-17071 Girona, Catalonia, Spain

 Supporting Information

ABSTRACT: Recently, much attention has been focused on the design and synthesis of molecules with aromatic Möbius topology. One of the most promising applications is the manufacture of Hückel-to-Möbius topological optical switches with high nonlinear optical properties. In this work, we evaluate the electronic and vibrational contributions to static and dynamic nonlinear optical properties of the C_S Hückel and C_2 Möbius topologies synthesized by Herges and co-workers (Ajami, D. et al. *Nature* **2003**, 426, 819). Calculations are performed at the HF, B3LYP, BHandHLYP, BMK, M052X, CAM-B3LYP, and MP2 levels with the 6-31+G(d) basis set. Our results conclude that the BHandHLYP, M052X, and CAM-B3LYP methods correctly reproduce the X-ray crystal structure and provide similar nonlinear optical properties.

1. INTRODUCTION

The stability of a molecular conformation is the result of different interactions, e.g., steric effect, presence of a number of hydrogen bonds, stacking interactions, and conjugation. When discussing the overall effect of these interactions, *aromaticity*¹ has often been a guiding principle. This concept has been one of the most controversial and debated issues in chemistry during the past two decades.² According to the well-known Hückel rule, an aromatic (hydrocarbon) compound is a cyclic planar molecular structure stabilized by the delocalization of $4n + 2 \pi$ electrons. On the other hand, if the planar molecule contains $4n \pi$ electrons, it is antiaromatic. However, the $4n + 2$ Hückel rule is not always fulfilled by the nonplanar molecules. In 1964, the seminal work of Heilbronner³ showed that the Hückel rule is reversed for π chemical systems with an odd number of half-twists. Heilbronner, based on Hückel molecular orbital theory, predicts that singlet annulenes with $4n \pi$ electrons would be aromatic systems in twisted conformations where the p orbitals lie on the surface of a Möbius strip, whereas singlet $[4n + 2]$ Möbius molecules would be antiaromatic.

The concept of Möbius aromaticity has become very popular in the literature.⁴ However, the synthesis of a viable⁵ aromatic Möbius system has been a challenge for 40 years. The difficulty of this synthesis resides in the curvature of the system, resulting in a destabilization due to the *ring strain* being larger than the stabilization due to the *Möbius aromaticity*. In order to reduce this ring strain, large cyclic molecules can be envisaged, although the structures become more flexible and can flip back to the less-strained and stable Hückel topology.

The first experimental evidence for an aromatic Möbius system was obtained in 1971 with a short-lived intermediate, the charged cyclic 8π electron [9]annulene, (C_9H_9)⁺.⁶ In 1998, Scheleyer et al.⁷ reported theoretical support of the aromaticity of this species using geometric and magnetic criteria. A recent study⁸ has shown that high-level coupled cluster calculations

predict that Hückel and Möbius isomers of (C_9H_9)⁺ are quasidegenerate (the energy difference is only 0.04 kcal mol⁻¹). Although experimental results⁸ indicate that the Hückel structure is the most stable conformation since it is the unique conformer detected, it was considered that the Möbius structure could exist in small concentrations below the detection limit.

In 2003, Herges and co-workers⁹ synthesized the first stable crystalline Möbius molecule combining a normal planar conjugated structure (with p orbitals orthogonal to the ring plane) and a rigid prefabricated belt-shaped conjugated segment (with p orbitals within the ring plane).^{9–11} The suprafacial–suprafacial allowed $[2 + 2]$ addition of syn-tricyclooctadiene and tetrahydrodianthracene followed by a thermal suprafacial–antarafacial cycloreversion leads to the isolation of five-ring opened isomers of bianthraquinodimethane modified [16]annulene. Two of them have Möbius topology (C_1 and C_2 symmetry, structure 2), and one of them shows a Hückel topology (C_S symmetry, structure 1, see Figure 1). The last two of them were isolated, but their structures could not be unequivocally elucidated. The aromaticity of these Möbius conformations has been a controversial issue, and it has been analyzed in different research groups,^{9,11,12} which concluded that they show a weak aromaticity.

The difficulty in synthesizing Möbius conformers with annulene systems arises from the small cis–trans isomerization barriers,¹³ and then alternative systems have been studied. For instance, the possibility of charged annulene anions and cations has been screened.¹⁴ Finally, the class of molecules that has succeeded in the creation of Möbius aromatic systems has been the expanded porphyrins,¹⁵ macrocycles of pyrrolic subunits, and their analogues (benzporphyrins, vacataporphyrins). The conformational flexibility, the number and the nature of substituents

Received: August 2, 2011

Published: October 14, 2011

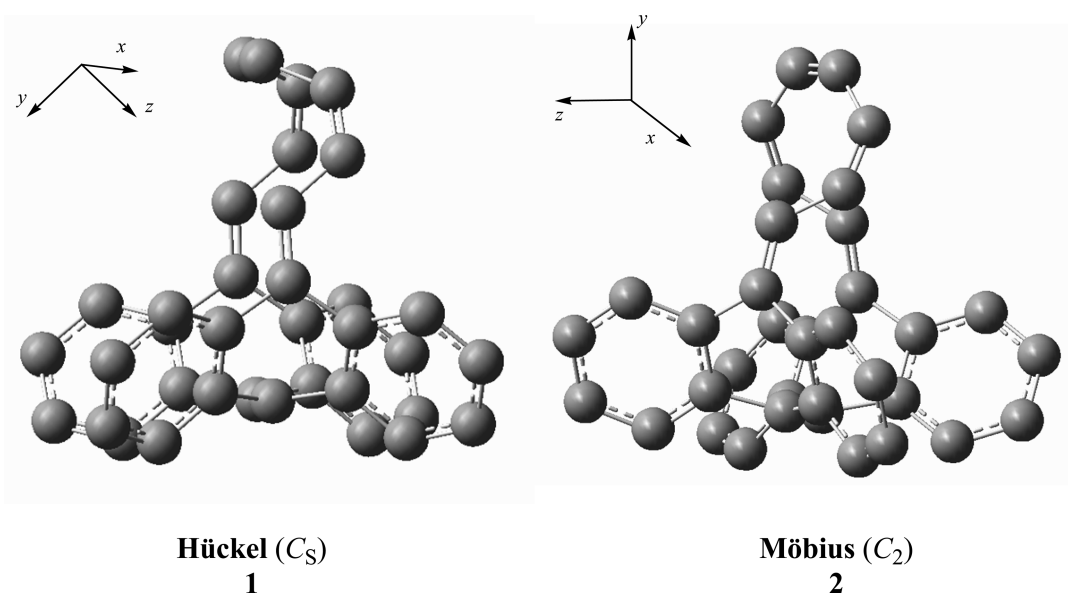


Figure 1. CAM-B3LYP/6-31+G(d) optimized geometries of C_5 Hückel and C_2 Möbius topologies of bianthraquinodimethane modified [16]annulene. The hydrogen atoms have been omitted for clarity.

on the pyrrolic and meso positions, and the metalation of the porphyrins allow them to achieve different topologies with distinct aromaticities and electronic properties.^{16–18}

One of the most appealing applications is the possibility of switching between Möbius and Hückel topologies, applying only small changes in the external conditions or in the structure of the ring. The Möbius and Hückel molecules may have different electronic structures and properties. Specifically, the expanded porphyrins present a clear relationship between aromaticity, molecular geometry, and nonlinear optical properties (NLOP).¹⁹ Then, topologically switchable porphyrins open the possibility of designing new optical switches. The key factors, which will determine their potential use as optical switches, will be the high values and the large differences of the NLOP between the Möbius and Hückel conformations.

The experimental results reported in the literature about NLOPs of Möbius–Hückel conformers show that they are very promising in the potential use as optical switches, although they only contain two-photon absorptions cross-section values. Computational chemistry is a useful tool for the evaluation of different static and dynamic NLO properties, e.g., the optical Kerr effect (OKE), electric-field-induced second harmonic generation (ESHG), and the intensity-dependent refractive index (IDRI). Xu et al.²⁰ studied the electronic dipole moment, polarizability, and first hyperpolarizability of four knot isomers of a cyclacene composed of 15 nitrogen-substituted benzo rings using the BHandHLYP/6-31+G(d) level. The four knot isomers are a Hückel cyclacene without a knot, a Möbius cyclacene with one knot, a Hückel cyclacene with two knots, and a Möbius cyclacene with three knots. The reported results show that the static electronic first polarizability values of cyclacenes with an odd number of knots (Möbius) can be 1 order of magnitude larger than those with an even number of knots (Hückel).

Using a clamped nucleus approximation,²¹ the polarizability and hyperpolarizabilities can be decomposed into electronic and vibrational contributions.²² Although neglected in the past, the vibrational contribution to the hyperpolarizabilities can be similar and even larger than the electronic counterpart.^{23–27}

The choice of a theoretical approach for the evaluation of NLOP is not an easy task. For instance, the desirable coupled cluster methods are computationally prohibitive for these Möbius–Hückel topological switches, and the MP2 could only be used for benchmarking purposes. Then, the only possible alternative method is density functional theory (DFT). However, it is well-known that an incorrect electric field dependence modeled by the exchange functional (in the conventional DFT methods) provokes an overestimation of the hyperpolarizabilities of chain-like molecules.²⁸ This deficiency can be alleviated using DFT functionals with a large fraction of Hartree–Fock (i.e., BMK, BHandHLYP, M052X)²⁹ or using new DFT functionals. Among them, it is important to remark on the long-range corrected functionals, which introduce a growing fraction of exact exchange when the distance increases, e.g., long-range corrected functional and the Coulomb-attenuating method (CAM).³⁰ These methods have become a promising tool for the evaluation of NLOP^{27,31–33} and chiroptical properties.³⁴

This work has two main goals: first, to report as far as we know the first exhaustive evaluation of the electronic and vibrational contributions to static and dynamic NLOP for a system with Hückel and Möbius topologies and, second, to search for a DFT level of theory capable of providing semiquantitative accuracy of the NLOP with a π -conjugated Hückel–Möbius switch. We have chosen the C_5 Hückel (structure 1, see Figure 2) and C_2 Möbius (structure 2) topologies of bianthraquinodimethane modified [16]annulene synthesized by Herges and co-workers.^{9,11} The moderate number of atoms of these systems allows a systematic study using different theoretical approaches.

2. COMPUTATIONAL METHODS

The evaluation of the electronic contribution to the dipole moment (μ^e), linear polarizability (α^e), first hyperpolarizability (β^e), and second hyperpolarizability (γ^e) were performed at the HF, B3LYP,³⁵ BHandHLYP,³⁶ BMK,³⁷ M052X,³⁸ CAM-B3LYP,³⁹ and MP2⁴⁰ levels with the 6-31+G(d)⁴¹ basis set using the Gaussian 09 program package.⁴² All of the NLOPs have been

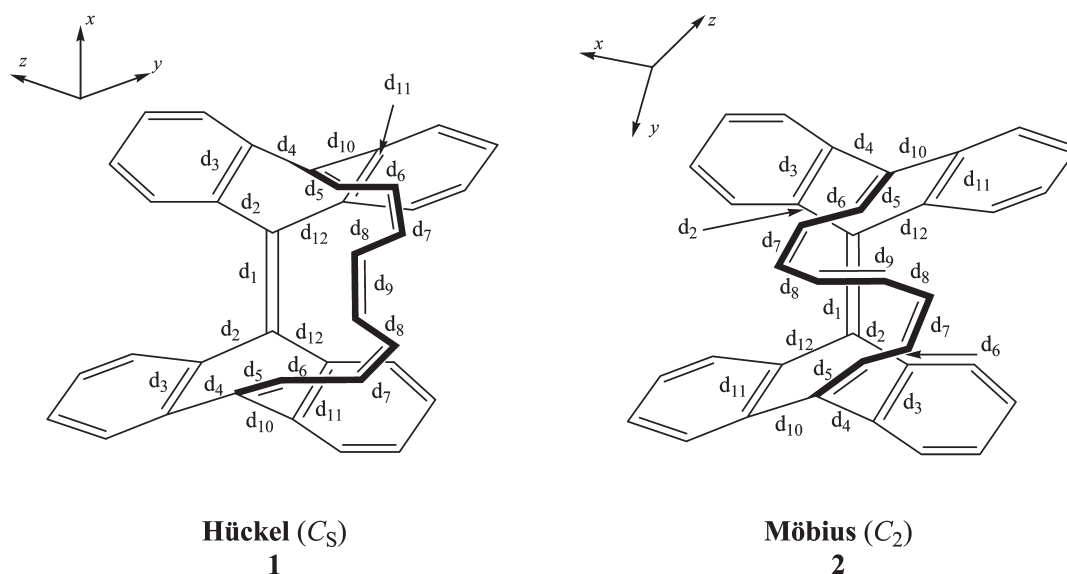


Figure 2. Representation of C_S Hückel and C_2 Möbius topologies of bianthraquinodimethane modified [16] annulene. The labels are the bond distances displayed in the Tables 1 and 2.

evaluated in the gas phase.⁴³ μ^e , α^e , and β^e were analytically evaluated for all of the methodologies, except for the MP2 level, the level at which only μ^e was obtained analytically due to computational limitations. All remaining high-order properties (i.e., γ^e for HF and DFT methods and α^e , β^e , and γ^e for the MP2 level) were obtained by finite field differentiation of the highest-order analytical electronic contribution available. The numerical differentiation was carried out for field strengths of ± 0.0002 , ± 0.0004 , ± 0.0008 , and ± 0.0016 au. The smallest field magnitude that produced a stable derivative was selected using a Romberg method triangle.⁴⁴ The symmetry restrictions have not been considered in the optimization process, and the structures of **1** and **2** at B3LYP/6-31G(d) obtained by Ajami et al.¹¹ have been used as the initial geometries of the optimization process. The average (hyper)polarizabilities are defined by following equations:²²

$$\bar{\alpha} = \frac{1}{3} \sum_{i=x,y,z} \alpha_{ii} \quad (1)$$

$$\bar{\beta} = \frac{1}{S|\bar{\mu}|} \sum_{i=x,y,z} \mu_i (\beta_{ijj} + \beta_{jij} + \beta_{jji}) \quad (2)$$

and

$$\bar{\gamma} = \frac{1}{15} \sum_{i,j=x,y,z} \gamma_{ijj} + \gamma_{jij} + \gamma_{jji} \quad (3)$$

The vibrational (hyper)polarizabilities can be computed using the pioneer perturbation treatment of Bishop and Kirtman⁴⁵ or the variational approach based on analytical response theory, proposed by Christiansen et al.⁴⁶ One approach intertwined with the BK method is the nuclear relaxation approach, whose derivation of vibrational NLOP formulas is based on determining the change equilibrium geometry induced by a static external field.^{47,48} Even though there is an exact correspondence between the BK perturbation treatment and the nuclear relaxation approach, the latter has spawned valuable new concepts and related computational procedures. From the viewpoint of the nuclear

relaxation (NR) procedure, it is natural to divide the total vibrational (hyper)polarizability into nuclear relaxation (P^{nr}) and curvature (P^c) contributions. P^{nr} and P^c arise from the change in the electronic and zero-point vibrational averaging corrections caused by the field-induced relaxation of the equilibrium geometry, respectively. The P^c is usually smaller and far more computationally expensive than P^{nr} ,^{24,26} and it is not computed here.

Under the infinite optical frequency (IOF) approximation, which corresponds to the limit $\omega \rightarrow \infty$, the expression for the dynamic P^{nr} can be obtained using the nuclear relaxation approach. Tests of the IOF approximation have shown that it yields satisfactory results.^{49,50} The bottleneck in calculating P^{nr} from analytical expressions is the number and the computational cost of the n th-order derivatives with respect to normal modes.⁴⁸ Their number is on the order of $(3N - 6)^n$, with N being the number of atoms. This problem can be circumvented by using the finite field nuclear relaxations approach or by introducing a set of static field-induced vibrational coordinates (FICs), which are just the displacement coordinates derived from the change in the equilibrium geometry induced by a static applied field.^{50–52} The FICs radically reduce the number of n th-order derivatives to be evaluated. For instance, for the nuclear relaxation contribution to Pockels effect, the analytical expressions containing sums over $3N - 6$ normal coordinates can be reduced to formulas that involve only three FICs.

The analytical definitions of the first (χ_1^α) and harmonic second-order ($\chi_{2,har}^{\alpha\beta}$) FICs are based on the expansion of the field-dependent displacement of the field-free normal coordinate (Q_i^F) induced by a uniform static electric field as a power series in the field (F_α). The expressions of χ_1^α and $\chi_{2,har}^{\alpha\beta}$ are given by⁵² nuclear coordinates

$$\chi_1^\alpha = \sum_{i=1}^{3N-6} \frac{\partial Q_i^F}{\partial F_\alpha} Q_i \quad (4)$$

and

$$\chi_{2,har}^{\alpha\beta} = \sum_{i=1}^{3N-6} \left(\frac{\partial^2 Q_i^F}{\partial F_\alpha \partial F_\beta} \right)_{har} Q_i \quad (5)$$

Table 1. Selected Geometrical Parameters of the Bianthraquinodimethane Modified [16]Annulene with Hückel Topology Calculated Using the 6-31+G(d) Basis Set and Seven Different Levels of Theory (All Quantities in Ångstroms)

distances ^a	HF	B3LYP	BHandHLYP	BMK	M052X	CAM-B3LYP	MP2
d_1	1.341	1.366	1.350	1.365	1.355	1.354	1.372
d_2	1.502	1.496	1.490	1.502	1.492	1.494	1.481
d_3	1.402	1.417	1.404	1.418	1.408	1.408	1.416
d_4	1.491	1.485	1.480	1.492	1.482	1.484	1.475
d_5	1.334	1.359	1.343	1.358	1.349	1.347	1.368
d_6	1.471	1.456	1.455	1.466	1.459	1.460	1.448
d_7	1.329	1.351	1.336	1.350	1.343	1.341	1.358
d_8	1.490	1.485	1.480	1.494	1.485	1.483	1.482
d_9	1.326	1.345	1.332	1.346	1.338	1.337	1.352
d_{10}	1.494	1.487	1.482	1.494	1.485	1.487	1.480
d_{11}	1.405	1.420	1.407	1.422	1.412	1.411	1.420
d_{12}	1.502	1.496	1.490	1.501	1.491	1.494	1.480
BLA	0.139	0.110	0.120	0.118	0.118	0.120	0.096

^a For the numbering of the distances, see Figure 2.**Table 2.** Selected Geometrical Parameters of the Bianthraquinodimethane Modified [16]Annulene with the Möbius Topology Calculated Using the 6-31+G(d) Basis Set and Seven Different Levels of Theory (All Quantities in Ångstroms)

distances ^a	HF	B3LYP	BHandHLYP	BMK	M052X	CAM-B3LYP	MP2	exptl ^b
d_1	1.338	1.363	1.347	1.362	1.353	1.352	1.373	1.353
d_2	1.497	1.491	1.485	1.496	1.486	1.489	1.475	1.488
d_3	1.400	1.416	1.402	1.417	1.406	1.407	1.415	1.411
d_4	1.494	1.486	1.482	1.493	1.483	1.486	1.474	1.485
d_5	1.336	1.364	1.346	1.362	1.353	1.351	1.373	1.355
d_6	1.462	1.445	1.446	1.456	1.450	1.450	1.436	1.449
d_7	1.334	1.362	1.344	1.360	1.351	1.349	1.370	1.350
d_8	1.475	1.458	1.459	1.469	1.463	1.463	1.454	1.452
d_9	1.336	1.364	1.346	1.362	1.354	1.351	1.374	1.332
d_{10}	1.493	1.486	1.481	1.493	1.484	1.485	1.475	1.485
d_{11}	1.407	1.421	1.408	1.423	1.413	1.412	1.422	1.412
d_{12}	1.503	1.497	1.491	1.503	1.493	1.495	1.483	1.489
BLA	0.130	0.094	0.108	0.103	0.105	0.107	0.077	0.104
SDD	0.014	0.012	0.007	0.014	0.008	0.007	0.018	

^a For the numbering of the distances, see Figure 2. ^b The experimental distances are obtained from the X-ray crystal structure, see ref 9.

Only one first order FIC is required to calculate each diagonal component of $\alpha^{\text{nr}}(0;0)$, $\beta^{\text{nr}}(0;0,0)$, $\beta^{\text{nr}}(\omega;\omega,0)_{\omega \rightarrow \infty}$ (IOF approximation NR Pockels β), and $\gamma^{\text{nr}}(-2\omega;\omega,\omega,0)_{\omega \rightarrow \infty}$ (IOF approximation NR field induced second harmonic (FISH) γ) tensors. Also, one unique FIC, but now the harmonic second order FIC, is necessary to calculate each diagonal component of $\gamma^{\text{nr}}(-\omega;\omega,-\omega,\omega)_{\omega \rightarrow \infty}$ (IOF approximation NR intensity-dependent refractive index (IDRI) γ). Using only two FICs, a first order and a harmonic second order FIC, one can also obtain each diagonal component of $\gamma^{\text{nr}}(-\omega;\omega,0,0)_{\omega \rightarrow \infty}$ (IOF approximation NR Kerr effect γ). The calculation of $\gamma^{\text{nr}}(0;0,0,0)$ requires second order FICs instead of harmonic second order FICs. The calculation of second order FICs is very expensive, and the static NR second hyperpolarizabilities have not been calculated here. The Supporting Information contains the plots and analysis of the no-mass-weighted orthonormal FICs of the Hückel and Möbius conformations evaluated at the CAM-B3LYP/6-31+G(d) level.

3. RESULTS AND DISCUSSION

Tables 1 and 2 show some selected bond lengths of the Hückel (structure 1) and Möbius (structure 2) topologies, respectively, using the 6-31+G(d) basis set and seven different levels of theory. The C–C bond lengths determined from X-ray crystallography are only reported for the Möbius structure.⁹ Tables 1 and 2 also contain the bond length alternation (BLA, the average difference between simple and double bond lengths) along the [16]annulene circuit of 1 and 2 structures. In both topologies, there are three possible [16]annulene paths. The BLA differences between the three paths are very small, less than 0.01 Å. Finally, Table 2 also includes the standard deviation of the distances (SDD), which is defined as

$$\text{SDD} = \sqrt{\sum_{i=1}^N \frac{(d_i - d_{i,\text{exp}})^2}{N-1}} \quad (6)$$

Table 3. Electronic Polarizabilities and First and Second Hyperpolarizabilities of the Bianthraquinodimethane Modified [16] Annulene with Hückel Topology Calculated Using the 6-31+G(d) Basis Set and Seven Different Levels of Theory (All Quantities in Atomic Units)^a

properties	HF	B3LYP	BHandHLYP	BMK	M052X	CAM-B3LYP	MP2
$\alpha_{xx}^e(0;0)$	3.68×10^2	3.90×10^2	3.71×10^2	3.75×10^2	3.65×10^2	3.74×10^2	3.58×10^2
$\alpha_{yy}^e(0;0)$	4.23×10^2	4.74×10^2	4.41×10^2	4.57×10^2	4.44×10^2	4.47×10^2	4.54×10^2
$\alpha_{zz}^e(0;0)$	4.00×10^2	4.45×10^2	4.18×10^2	4.34×10^2	4.23×10^2	4.24×10^2	4.30×10^2
$\bar{\alpha}^e(0;0)$	3.97×10^2	4.36×10^2	4.10×10^2	4.22×10^2	4.10×10^2	4.15×10^2	4.14×10^2
$\beta_{yyyy}^e(0;0,0)$	9.13×10^1	2.52×10^2	1.57×10^2	1.41×10^2	1.07×10^2	1.39×10^2	1.69×10^2
$\beta_{zzzz}^e(0;0,0)$	4.90×10^1	1.80×10^2	1.07×10^2	9.22×10^1	6.90×10^1	8.97×10^1	1.08×10^2
$\bar{\beta}^e(0;0,0)$	-9.71×10^1	-2.75×10^2	-1.78×10^2	-1.88×10^2	-1.46×10^2	-1.67×10^2	-2.16×10^2
$\gamma_{xxxx}^e(0;0,0,0)$	7.59×10^4	1.82×10^5	1.15×10^5	1.38×10^5	1.18×10^5	1.20×10^5	1.53×10^5
$\gamma_{yyyy}^e(0;0,0,0)$	1.36×10^5	3.08×10^5	2.10×10^5	2.22×10^5	2.11×10^5	2.17×10^5	2.76×10^5
$\gamma_{zzzz}^e(0;0,0,0)$	9.00×10^4	2.09×10^5	1.38×10^5	1.39×10^5	1.38×10^5	1.43×10^5	1.74×10^5
$\bar{\gamma}^e(0;0,0,0)$	3.38×10^4	7.81×10^4	5.14×10^4	5.44×10^4	5.13×10^4	5.35×10^4	

^a $\beta_{xxx}^e(0;0,0)$ is null by symmetry.

Table 4. Electronic Polarizabilities and First and Second Hyperpolarizabilities of the Bianthraquinodimethane Modified [16] Annulene with the Möbius Topology Calculated Using the 6-31+G(d) Basis Set and Seven Different Levels of Theory (All Quantities in Atomic Units)^a

properties	HF	B3LYP	BHandHLYP	BMK	M052X	CAM-B3LYP	MP2
$\alpha_{xx}^e(0;0)$	3.89×10^2	4.52×10^2	4.18×10^2	4.43×10^2	4.32×10^2	4.26×10^2	4.52×10^2
$\alpha_{yy}^e(0;0)$	4.39×10^2	4.93×10^2	4.58×10^2	4.73×10^2	4.60×10^2	4.63×10^2	4.62×10^2
$\alpha_{zz}^e(0;0)$	3.71×10^2	3.91×10^2	3.69×10^2	3.70×10^2	3.54×10^2	3.72×10^2	3.49×10^2
$\bar{\alpha}^e(0;0)$	4.00×10^2	4.45×10^2	4.15×10^2	4.29×10^2	4.15×10^2	4.20×10^2	4.21×10^2
$\beta_{yyyy}^e(0;0,0)$	-8.25×10^1	-2.75×10^2	-1.92×10^2	-2.61×10^2	-2.50×10^2	-2.22×10^2	-5.31×10^2
$\bar{\beta}^e(0;0,0)$	9.26×10^1	3.40×10^2	2.21×10^2	-3.37×10^2	-2.71×10^2	2.41×10^2	-5.13×10^2
$\gamma_{xxxx}^e(0;0,0,0)$	9.15×10^4	2.27×10^5	1.43×10^5	1.45×10^5	1.38×10^5	1.48×10^5	1.75×10^5
$\gamma_{yyyy}^e(0;0,0,0)$	9.57×10^4	2.45×10^5	1.62×10^5	1.66×10^5	1.68×10^5	1.75×10^5	2.67×10^5
$\gamma_{zzzz}^e(0;0,0,0)$	9.59×10^4	2.23×10^5	1.40×10^5	1.37×10^5	1.31×10^5	1.42×10^5	1.65×10^5
$\bar{\gamma}^e(0;0,0,0)$	3.88×10^4	1.00×10^5	6.27×10^4	6.95×10^4	6.77×10^4	6.57×10^4	

^a $\beta_{xxx}^e(0;0,0)$ and $\beta_{zzz}^e(0;0,0)$ are null by symmetry.

where d_i and $d_{i,\text{exp}}$ are the calculated and experimental bond distances, respectively, and N is the number of bond lengths considered ($N = 12$). From the BLA and SDD values, one can easily see that a better geometrical description of the [16]annulene path is obtained with the BHandHLYP, M052X, and CAM-B3LYP methodologies than using HF, B3LYP, BMK, and MP2 treatments.

Moreover, Tables 1 and 2 show that HF is the method that predicts the shortest double bonds and the longest single bonds along the [16]annulene circuit, i.e., the largest BLA value. On the other hand, MP2 is the method that predicts the longest double bonds and the shortest single bonds, i.e., the smallest BLA value. The bond distances calculated from the B3LYP and BMK approaches are larger than the experimental ones. It is worth noting that the BHandHLYP and CAM-B3LYP methods show the smallest values of SDD, although BHandHLYP predicts a slightly large BLA. Finally, the M052X treatment obtains the closest BLA distance to the experimental value.

According to the results displayed in Table 2, one can conclude that M052X, BHandHLYP, and CAM-B3LYP levels of theory provide an accurate reproduction of the X-ray crystal structure of the bianthraquinodimethane modified [16]annulene with Möbius topology.

Tables 3 and 4 contain the electronic contribution to α , β , and γ for the structures **1** and **2**, respectively, using the 6-31+G(d)

basis set and seven different levels of theory. For each property, the diagonal components and the average values, see eqs 1–3, are reported. As one can easily see, the first hyperpolarizabilities of these two conformations are quite small, and no relevant conclusions can be obtained. In addition, the terms of the dipole moment, which have not been reported in the Tables 3 and 4, are also negligible (it is important to remember that conformations **1** and **2** show C_s and C_2 symmetry, respectively, and some diagonal terms are null by symmetry). Then, in the following paragraphs our analysis will focus on α and γ values. The evaluation of $\bar{\gamma}^e$ has not been reported at the MP2 level due to computational limitations; i.e., $\bar{\gamma}^e$ requires terms like γ_{ijij}^e , see eq 3, which are not easy to evaluate from the finite field differentiation of μ^e .

In Tables 3 and 4, one can notice the electronic correlation results essential for the correct evaluation of the NLOP. For instance, the MP2 and CAM-B3LYP methodologies increase γ^e values around 100% and 60% with respect to the HF level (around 10% for α^e values). The MP2 results show larger α^e and γ^e values than HF, BHandHLYP, BMK, M052X, and CAM-B3LYP methods and smaller values than B3LYP. Although in the literature it has been reported that MP2 treatment often yields a significant fraction of the electron correlation contribution of the NLOP,^{25,26} it has also been shown that the MP2 approach overestimates by more than a factor of 2 with respect to CCSD(T) for

Table 5. Nuclear Relaxation Polarizabilities and First Hyperpolarizabilities of the Bianthraquinodimethane Modified [16] Annulene with the Hückel and Möbius Topologies Calculated at HF and CAM-B3LYP Levels of Theory Using the 6-31+G(d) Basis Set (All Quantities in Atomic Units)^a

	HF		CAM-B3LYP	
	Hückel	Möbius	Hückel	Möbius
$\alpha_{xx}^{\text{nr}}(0;0)$	2.70×10^1	1.68×10^1	3.19×10^1	1.20×10^1
$\alpha_{yy}^{\text{nr}}(0;0)$	1.34×10^1	9.92×10^0	1.31×10^1	9.30×10^0
$\alpha_{zz}^{\text{nr}}(0;0)$	1.01×10^1	1.69×10^1	1.08×10^1	2.18×10^1
$\bar{\alpha}^{\text{nr}}(0;0)$	1.69×10^1	1.46×10^1	1.86×10^1	1.44×10^1
$\beta_{yyy}^{\text{nr}}(0;0,0)$	-1.30×10^2	-2.63×10^2	-2.10×10^2	-2.55×10^2
$\beta_{zzz}^{\text{nr}}(0;0,0)$	-4.04×10^2		-5.51×10^2	
$\bar{\beta}^{\text{nr}}(0;0,0)$	3.48×10^2	5.83×10^2	1.92×10^3	7.45×10^2
$\beta_{yyy}^{\text{nr}}(-\omega;\omega,0)_{\omega \rightarrow \infty}$	-6.96×10^1	-8.16×10^1	-9.06×10^1	-8.38×10^1
$\beta_{zzz}^{\text{nr}}(-\omega;\omega,0)_{\omega \rightarrow \infty}$	-1.63×10^2		-2.13×10^2	
$\bar{\beta}^{\text{nr}}(-\omega;\omega,0)_{\omega \rightarrow \infty}$	1.24×10^2	2.07×10^2	6.11×10^2	2.62×10^2

^a $\beta_{xxx}^{\text{nr}}(0;0,0)$ and $\beta_{xxx}^{\text{nr}}(-\omega;\omega,0)_{\omega \rightarrow \infty}$ are null by symmetry, and $\beta_{zzz}^{\text{nr}}(0;0,0)$ and $\beta_{zzz}^{\text{nr}}(-\omega;\omega,0)_{\omega \rightarrow \infty}$ are also null by symmetry, but only at the Möbius conformation.

Table 6. Nuclear Relaxation Second Hyperpolarizabilities of the Bianthraquinodimethane Modified [16] Annulene with the Hückel and Möbius Topologies Calculated at HF and CAM-B3LYP Levels of Theory Using the 6-31+G(d) Basis Set (All Quantities in Atomic Units)

	HF		CAM-B3LYP	
	Hückel	Möbius	Hückel	Möbius
$\gamma_{xxxx}^{\text{nr}}(-\omega;\omega,0,0)_{\omega \rightarrow \infty}$	1.00×10^5	3.48×10^4	1.74×10^5	4.05×10^4
$\gamma_{yyyy}^{\text{nr}}(-\omega;\omega,0,0)_{\omega \rightarrow \infty}$	5.58×10^4	8.11×10^4	6.69×10^4	1.08×10^5
$\gamma_{zzzz}^{\text{nr}}(-\omega;\omega,0,0)_{\omega \rightarrow \infty}$	4.46×10^4	3.07×10^4	6.23×10^4	6.45×10^4
$\bar{\gamma}^{\text{nr}}(-\omega;\omega,0,0)_{\omega \rightarrow \infty}$ ^{a,b}	1.16×10^5	1.06×10^5	6.07×10^5	1.44×10^5
$\gamma_{xxxx}^{\text{nr}}(-2\omega;\omega,\omega,0)_{\omega \rightarrow \infty}$	3.92×10^3	1.86×10^3	4.93×10^3	1.38×10^3
$\gamma_{yyyy}^{\text{nr}}(-2\omega;\omega,\omega,0)_{\omega \rightarrow \infty}$	2.11×10^3	6.67×10^2	1.56×10^3	3.44×10^2
$\gamma_{zzzz}^{\text{nr}}(-2\omega;\omega,\omega,0)_{\omega \rightarrow \infty}$	1.68×10^3	2.02×10^3	1.75×10^3	2.98×10^3
$\bar{\gamma}^{\text{nr}}(-2\omega;\omega,\omega,0)_{\omega \rightarrow \infty}$	2.73×10^3	2.04×10^3	2.48×10^3	2.02×10^3
$\gamma_{xxxx}^{\text{nr}}(-\omega;\omega,-\omega,\omega)_{\omega \rightarrow \infty}$	1.75×10^5	6.29×10^4	2.41×10^5	7.89×10^4
$\gamma_{yyyy}^{\text{nr}}(-\omega;\omega,-\omega,\omega)_{\omega \rightarrow \infty}$	9.94×10^4	1.61×10^5	1.22×10^5	2.19×10^5
$\gamma_{zzzz}^{\text{nr}}(-\omega;\omega,-\omega,\omega)_{\omega \rightarrow \infty}$	8.39×10^4	5.52×10^4	1.19×10^5	1.20×10^5
$\bar{\gamma}^{\text{nr}}(-\omega;\omega,-\omega,\omega)_{\omega \rightarrow \infty}$ ^a	2.17×10^5	1.21×10^5	5.49×10^5	1.68×10^5

^a The $[\alpha^2]$ terms were calculated using $3N - 6$ normal modes instead of the six first (χ_1^{α}) and harmonic second-order ($\chi_{2,\text{har}}^{\alpha\alpha}$) FICs in order to obtain the correct value for the nondiagonal elements of the hyperpolarizability tensor (see Table 1 of ref 52 for more details).^b The nondiagonal elements of the $[\mu^2\alpha]$ term are approximate. The exact calculation of such elements requires using either second-order $\chi_{2,\text{har}}^{xy}$, $\chi_{2,\text{har}}^{xz}$, and $\chi_{2,\text{har}}^{yz}$ FICs or the $3N - 6$ normal coordinates (see Table 1 of ref 52 for more details).

polybutatriene chains.⁵³ In the previous paragraphs, we have shown that MP2 methodology predicts a Möbius geometry with the largest SDD value (0.018 Å) with respect to the X-ray crystal structure. Then, one can conclude that for these systems the MP2 level cannot be considered as a reference model for the evaluation of the NLOP.

Among the seven methods used in this work, the HF and B3LYP methods present the smallest and largest, respectively, values of α^e and γ^e . The overestimation of the second hyperpolarizabilities using the B3LYP approach is expected due to the incorrect electric field dependence modeled by the conventional exchange functional treatments.²⁸ On the other hand, the four remaining DFT methods used in this work (BHandHLYP, BMK, M052X, and CAM-B3LYP) show intermediate values of α^e and γ^e between HF and B3LYP levels. It is worth noting that α^e and

γ^e values obtained with these four DFT treatments are very similar; i.e., the differences between them are always smaller than 20%. It is important to remember that BHandHLYP, M052X, and CAM-B3LYP levels correctly reproduce the X-ray crystal structure of bianthraquinodimethane modified [16]annulene with Möbius topology. To validate the NLOP results obtained using these four DFT methodologies, it would be necessary to evaluate them using coupled cluster methods results, although they are computationally prohibitive for **1** and **2** structures. Nevertheless, several works^{32,33} have shown that DFT functionals with a large fraction of Hartree–Fock and DFT long-range functionals remove the overestimation of polarizabilities and hyperpolarizabilities by standard DFT. Moreover, it has been found that CAM-B3LYP results are very similar to the desirable coupled cluster methods.³³ We conclude that the results of this

work are encouraging to the evaluation of the NLOP for larger Hückel–Möbius switches, because it seems that BHandHLYP, M052X, and CAM-B3LYP methods can provide semiquantitative accuracy with a reasonable computational cost.

The diagonal components of α and γ for the Möbius and Hückel topologies are quite similar. Only the y direction (more or less the direction that goes through the [16]annulene circuit) shows slightly higher values than x and z directions; e.g., in the Hückel conformation, $\gamma_{yyyy}^e(0;0,0,0)$ is between 35 and 45% larger than $\gamma_{xxxx}^e(0;0,0,0)$ and $\gamma_{zzzz}^e(0;0,0,0)$ for the seven different levels of theory, see Table 3. Moreover, it is worth noting that the $\bar{\alpha}^e$ and $\bar{\gamma}^e$ values for the Hückel and Möbius conformations are also very similar between them. The maximum difference is obtained at the M052X level with a $\bar{\gamma}^e(0;0,0,0)_{\text{Hückel}}/\bar{\gamma}^e(0;0,0,0)_{\text{Möbius}}$ ratio of 0.76. This conclusion shows a direct link with the controversial issue about the aromaticity character of these systems. Our calculations agree well with previous results^{9,11,12} that **1** and **2** structures present similar electronic and magnetic properties.

In Tables 5 and 6 are reported the static and dynamic nuclear relaxation polarizabilities and first and second hyperpolarizabilities for structures **1** and **2**. Our analysis has been focused on only two methodologies (HF and CAM-B3LYP), because the evaluation of the vibrational contribution to NLOP implies an important computational effort (each conformation requires 13 frequency calculations). These two treatments have been selected to check the differences between the HF level and a theoretical approach that correctly describes the geometry and the electronic contribution to NLOP. The CAM-B3LYP method was chosen, although as we have seen in previous paragraphs M052X and BHandHLYP could also have been good choices. In a similar way to the electronic contribution, CAM-B3LYP treatment gives important augmentations for $\bar{\gamma}^{\text{nr}}$ with respect to the values obtained at the HF level, e.g., $\bar{\gamma}^{\text{nr}}(-\omega;\omega,0,0)_{\omega \rightarrow \infty}$ and $\bar{\gamma}^{\text{nr}}(-\omega;\omega,-\omega,\omega)_{\omega \rightarrow \infty}$ at the Hückel conformation increase around 400% and 150%, respectively.

In contrast to Tables 3 and 4, important differences are found between the diagonal components of nuclear relaxation α and γ for Hückel and Möbius topologies. For instance, at the CAM-B3LYP level in the Möbius conformation, $\alpha_{zz}^{\text{nr}}(0;0)$ is 1.8 and 2.3 times larger than $\alpha_{xx}^{\text{nr}}(0;0)$ and $\alpha_{yy}^{\text{nr}}(0;0)$, respectively (see Table 5), and $\gamma_{yyyy}^{\text{nr}}(-\omega;\omega,0,0)_{\omega \rightarrow \infty}$ is 2.7 and 1.7 times larger than in the x and z directions, respectively (see Table 6). In analogy to the electronic contribution, $\bar{\alpha}^{\text{nr}}$ and $\bar{\gamma}^{\text{nr}}$ values for the Hückel and Möbius conformations are quite similar between them using the HF treatment; i.e., the maximum difference is the ratio $\bar{\gamma}^{\text{nr}}(-2\omega;\omega,\omega,0)_{\text{Hückel}}/\bar{\gamma}^{\text{nr}}(-2\omega;\omega,\omega,0)_{\text{Möbius}}$ with a value of 1.8. However, CAM-B3LYP results show two important exceptions of this tendency, the average IOF approximation NR Kerr and IDRI effects, which present ratios of 4.2 and 3.3, respectively.

Last but not least, we analyze the relevance of the vibrational contribution to NLOP with respect to the electronic contribution. According to Tables 3–6, it is clear that $\bar{\alpha}^{\text{nr}}(0;0)$ and $\bar{\gamma}^{\text{nr}}(-2\omega;\omega,\omega,0)_{\omega \rightarrow \infty}$ corrections are not negligible, though they are less than 10% of the corresponding static electronic properties. Moreover, $\bar{\gamma}^{\text{nr}}(-\omega;\omega,0,0)_{\omega \rightarrow \infty}$ and $\bar{\gamma}^{\text{nr}}(-\omega;\omega,-\omega,\omega)_{\omega \rightarrow \infty}$ are either larger than or comparable in size to $\bar{\gamma}^e(0;0,0,0)$. For instance, the IOF approximation of NR IDRI averages of structures **1** and **2** evaluated at the CAM-B3LYP level are 10.2 and 2.6, respectively, larger than the electronic contribution. Then, one can conclude that an accurate evaluation of NLOP for annulenes with Hückel and Möbius

topologies requires the study of the vibrational contribution. In addition, the sum of the electronic and vibrational contributions ($\bar{\gamma} = \bar{\gamma}^e + \bar{\gamma}^{\text{vib}}$) shows that the structures **1** and **2** present high values of NLOP, i.e., values around 5×10^5 .

4. SUMMARY AND CONCLUSIONS

We have investigated the electronic and vibrational contributions to static and dynamic NLOP of the bianthraquinodimethane modified [16]annulene with C_5 Hückel (structure **1**) and C_2 Möbius (structure **2**) topologies synthesized by Herges and co-workers. The calculations were performed at the HF, B3LYP, BHandHLYP, BMK, M052X, CAM-B3LYP, and MP2 levels with the 6-31+G(d) basis set. No analogous treatments of NLOP for these systems have been carried out as far as we know. The results of this work lead us to the following conclusions:

- Among the seven treatments considered in this work, the BHandHLYP, M052X, and CAM-B3LYP levels correctly reproduce the X-ray crystal structure of the annulene with Möbius topology. On the contrary, the HF, MP2, and B3LYP methods predict structural conformations, which show important divergences with respect to the experimental results.
- In the study of the electronic contribution to NLOP, it has been found that BHandHLYP, M052X, and CAM-B3LYP methods provide similar results. On the other hand, HF (MP2 and B3LYP) underestimates (overestimate) the NLOP. In the literature, it has been reported that CAM-B3LYP agrees with CCSD(T). Then, we conclude that BHandHLYP, M052X, and CAM-B3LYP can provide semiquantitative results with a reasonable computational cost for the evaluation of NLOP with Hückel–Möbius switches.
- $\bar{\alpha}$ and $\bar{\gamma}$ values for Hückel and Möbius conformations are similar. Our results for the NLOP agree with literature results that both Hückel and Möbius structures have similar electronic and magnetic properties.
- The vibrational contribution to static and dynamic NLOP for Möbius–Hückel systems can be either larger or comparable in size than the electronic contribution, and it must be considered for an accurate evaluation of the NLOP.

The experience obtained in this work will be very useful for the photophysical characterization of new topologically switchable porphyrins with high NLOPs (see refs 15–18), which show a clear relationship between aromaticity, molecular geometry, and NLOP. The key factors, which will determine their potential as optical switches, will be the high values and the large differences of NLOP between the Möbius and Hückel conformations. These systems present a considerable number of atoms (around 150 atoms), and therefore, the choice of an adequate methodology for the electronic and vibrational contributions is essential for the correct prediction of the NLOP. Additional work on the evaluation of new topologically switchable porphyrins (A,D-di-*p*-benz[28]hexaphyrin(1.1.1.1.1.1)¹⁶ and meso-aryl-substituted [28]hexaphyrins(1.1.1.1.1.1)¹⁸) with high NLOP is in progress in our laboratory.

■ ASSOCIATED CONTENT

S Supporting Information. Plots and analysis of χ_1^α and $\chi_{2,\text{har}}^{\alpha\alpha}$ of the Hückel and Möbius conformations evaluated at the CAM-B3LYP/6-31+G(d) level studied in this work.

This material is available free of charge via the Internet at <http://pubs.acs.org>.

AUTHOR INFORMATION

Corresponding Author

*Phone: +34 934006111. Fax: +34 932045903. E-mail: mtsqbm@iqac.csic.es.

ACKNOWLEDGMENT

This research has been supported by the Spanish Dirección General de Investigación Científica y Técnica (DGYCIT, grant CTQ2008-06536/BQU), the Generalitat de Catalunya (Grant 2009SGR01472), and the Research Executive Agency (Grant Agreement no. PERG05-GA-2009-249310). The calculations described in this work were carried out at the Centre de Supercomputació de Catalunya (CESCA). M.T.-S. acknowledges the CSIC for the JAE-DOC contract.

REFERENCES

- (1) Faraday, M. *Trans. R. Soc. London* **1825**, *105*, 440.
- (2) Minkin, V. I.; Glukhovtsev, M. N.; Simkin, B. Y. *Aromaticity and Antiaromaticity*; John Wiley & Sons: New York, 1994. Schleyer, P. v. R. *Chem. Rev.* **2001**, *101*, 1115. De Prof, F.; Geerlings, P. *Chem. Rev.* **2001**, *101*, 1451. Chen, Z. F.; Wannere, C. S.; Corminboeuf, C.; Puchta, R.; Schleyer, P. V. *Chem. Rev.* **2005**, *105*, 3842. Poater, J.; Duran, M.; Solà, M.; Silvi, B. *Chem. Rev.* **2005**, *105*, 3911.
- (3) Heilbronner, E. *Tetrahedron Lett.* **1964**, 1923.
- (4) Zimmerman, H. E. *J. Am. Chem. Soc.* **1966**, *88*, 1564. Castro, C.; Karney, W. L.; Valencia, M. A.; Vu, C. M. H.; Pemberton, R. P. *J. Am. Chem. Soc.* **2005**, *127*, 9704. Rzepa, H. S. *Chem. Rev.* **2005**, *105*, 3697. Pemberton, R. P.; McShane, C. M.; Castro, C.; Karney, W. L. *J. Am. Chem. Soc.* **2006**, *128*, 16692. Moll, J. F.; Pemberton, R. P.; Gutierrez, M. G.; Castro, C.; Karney, W. L. *J. Am. Chem. Soc.* **2007**, *129*, 274. Allan, C. S. M.; Rzepa, H. S. *J. Chem. Theory Comput.* **2008**, *4*, 1841. Rappaport, S. M.; Rzepa, H. S. *J. Am. Chem. Soc.* **2008**, *130*, 7613. Rzepa, H. S. *Inorg. Chem.* **2008**, *47*, 8932. Mauksch, M.; Tsogoeva, S. B. *Chem.—Eur. J.* **2010**, *16*, 7843.
- (5) Hoffmann, R.; Schleyer, P. v. R.; Schaefer, H. F. *Angew. Chem., Int. Ed.* **2008**, *47*, 7164.
- (6) Barborak, J. C.; Su, T. M.; Schleyer, P. v. R.; Boche, G.; Schneide, G. *J. Am. Chem. Soc.* **1971**, *93*, 279. Anastass, A. G.; Yakali, E. *J. Am. Chem. Soc.* **1971**, *93*, 3803. Anastass, A. G.; Eachus, S. W.; Yakali, E.; Elliott, R. L. *J. Chem. Soc. Chem. Commun.* **1972**, 531.
- (7) Mauksch, M.; Gogonea, V.; Jiao, H.; Schleyer, P. v. R. *Angew. Chem., Int. Ed.* **1998**, *37*, 2395.
- (8) Bucher, G.; Grimme, S.; Huenerbein, R.; Auer, A. A.; Mucke, E.; Kohler, F.; Siegwirth, J.; Herges, R. *Angew. Chem., Int. Ed.* **2009**, *48*, 9971.
- (9) Ajami, D.; Oeckler, O.; Simon, A.; Herges, R. *Nature* **2003**, *426*, 819.
- (10) Herges, R. *Chem. Rev.* **2006**, *106*, 4820.
- (11) Ajami, D.; Hess, K.; Kohler, F.; Nather, C.; Oeckler, O.; Simon, A.; Yamamoto, C.; Okamoto, Y.; Herges, R. *Chem.—Eur. J.* **2006**, *12*, 5434.
- (12) Castro, C.; Chen, Z. F.; Wannere, C. S.; Jiao, H. J.; Karney, W. L.; Mauksch, M.; Puchta, R.; Hommes, N.; Schleyer, P. v. R. *J. Am. Chem. Soc.* **2005**, *127*, 2425.
- (13) Mohebbi, A. R.; Mucke, E. K.; Schaller, G. R.; Kohler, F.; Sonnichsen, F. D.; Ernst, L.; Nather, C.; Herges, R. *Chem.—Eur. J.* **2010**, *16*, 7767.
- (14) Warner, P. M. *J. Org. Chem.* **2006**, *71*, 9271. Mauksch, M.; Tsogoeva, S. B. *Eur. J. Org. Chem.* **2008**, 5755. Wannere, C. S.; Rzepa, H. S.; Rinderspacher, B. C.; Paul, A.; Allan, C. S. M.; Schaefer, H. F.; Schleyer, P. v. R. *J. Phys. Chem. A* **2009**, *113*, 11619. Mucke, E. K.; Kohler, F.; Herges, R. *Org. Lett.* **2010**, *12*, 1708.
- (15) Jux, N. *Angew. Chem., Int. Ed.* **2008**, *47*, 2543. Yoon, Z. S.; Osuka, A.; Kim, D. *Nature Chem.* **2009**, *1*, 113. Shin, J. Y.; Kim, K. S.; Yoon, M. C.; Lim, J. M.; Yoon, Z. S.; Osuka, A.; Kim, D. *Chem. Rev.* **2010**, *39*, 2751. Osuka, A.; Saito, S. *Chem. Commun.* **2011**, *47*, 4330. Saito, S.; Osuka, A. *Angew. Chem., Int. Ed.* **2011**, *50*, 4342. Stepień, M.; Sprutta, N.; Latos-Grażyński, L. *Angew. Chem., Int. Ed.* **2011**, *50*, 4288.
- (16) Stepień, M.; Latos-Grażyński, L.; Sprutta, N.; Chwalisz, P.; Sztierenberg, L. *Angew. Chem., Int. Ed.* **2007**, *46*, 7869.
- (17) Saito, S.; Shin, J. Y.; Lim, J. M.; Kim, K. S.; Kim, D.; Osuka, A. *Angew. Chem., Int. Ed.* **2008**, *47*, 9657. Tanaka, Y.; Saito, S.; Mori, S.; Aratani, N.; Shinokubo, H.; Shibata, N.; Higuchi, Y.; Yoon, Z. S.; Kim, K. S.; Noh, S. B.; Park, J. K.; Kim, D.; Osuka, A. *Angew. Chem., Int. Ed.* **2008**, *47*, 681. Park, J. K.; Yoon, Z. S.; Yoon, M. C.; Kim, K. S.; Mori, S.; Shin, J. Y.; Osuka, A.; Kim, D. *J. Am. Chem. Soc.* **2008**, *130*, 1824. Shin, J. Y.; Lim, J. M.; Yoon, Z. S.; Kim, K. S.; Yoon, M. C.; Hiroto, S.; Shinokubo, H.; Shimizu, S.; Osuka, A.; Kim, D. *J. Phys. Chem. B* **2009**, *113*, 5794. Kim, K. S.; Yoon, Z. S.; Ricks, A. B.; Shin, J. Y.; Mori, S.; Sankar, J.; Saito, S.; Jung, Y. M.; Wasielewski, M. R.; Suka, A.; Kim, D. *J. Phys. Chem. A* **2009**, *113*, 4498. Yoon, M. C.; Cho, S.; Suzuki, M.; Osuka, A.; Kim, D. *J. Am. Chem. Soc.* **2009**, *131*, 7360. Tokuiji, S.; Shin, J. Y.; Kim, K. S.; Lim, J. M.; Youfu, K.; Saito, S.; Kim, D.; Osuka, A. *J. Am. Chem. Soc.* **2009**, *131*, 7240. Inoue, M.; Kim, K. S.; Suzuki, M.; Lim, J. M.; Shin, J. Y.; Kim, D.; Osuka, A. *Angew. Chem., Int. Ed.* **2009**, *48*, 6687. Lim, J. M.; Shin, J. Y.; Tanaka, Y.; Saito, S.; Osuka, A.; Kim, D. *J. Am. Chem. Soc.* **2010**, *132*, 3105. Stepień, M.; Szyzsko, B.; Latos-Grażyński, L. *J. Am. Chem. Soc.* **2010**, *132*, 3140. Tanaka, T.; Sugita, T.; Tokuiji, S.; Saito, S.; Osuka, A. *Angew. Chem., Int. Ed.* **2010**, *49*, 6619. Higashino, T.; Lim, J. M.; Miura, T.; Saito, S.; Shin, J. Y.; Kim, D.; Osuka, A. *Angew. Chem., Int. Ed.* **2010**, *49*, 4950. Yoon, M. C.; Shin, J. Y.; Lim, J. M.; Saito, S.; Yoneda, T.; Osuka, A.; Kim, D. *Chem.—Eur. J.* **2011**, *17*, 6707.
- (18) Sankar, J.; Mori, S.; Saito, S.; Rath, H.; Suzuki, M.; Inokuma, Y.; Shinokubo, H.; Kim, K. S.; Yoon, Z. S.; Shin, J. Y.; Lim, J. M.; Matsuzaki, Y.; Matsushita, O.; Muranaka, A.; Kobayashi, N.; Kim, D.; Osuka, A. *J. Am. Chem. Soc.* **2008**, *130*, 13568.
- (19) Lim, J. M.; Yoon, Z. S.; Shin, J. Y.; Kim, K. S.; Yoon, M. C.; Kim, D. *Chem. Commun.* **2009**, 261.
- (20) Xu, H. L.; Li, Z. R.; Su, Z. M.; Muhammad, S.; Gu, F. L.; Harigaya, K. *J. Phys. Chem. C* **2009**, *113*, 15380.
- (21) Bishop, D. M.; Kirtman, B.; Champagne, B. *J. Chem. Phys.* **1997**, *107*, 5780.
- (22) Bishop, D. M.; Norman, P. In *Handbook of Advanced Electronic and Photonic Materials*; Nalwas, H. S., Ed.; Academic: San Diego, 2001; Vol. 9, p 1.
- (23) Champagne, B.; Luis, J. M.; Duran, M.; Andres, J. L.; Kirtman, B. *J. Chem. Phys.* **2000**, *112*, 1011. Kirtman, B.; Champagne, B.; Luis, J. M. *J. Comput. Chem.* **2000**, *21*, 1572. Milledori, S.; Alparone, A. *Phys. Chem. Chem. Phys.* **2000**, *2*, 2495. Champagne, B.; Spassova, M.; Jadin, J. B.; Kirtman, B. *J. Chem. Phys.* **2002**, *116*, 3935. Torrent-Sucarrat, M.; Luis, J. M.; Kirtman, B. *J. Chem. Phys.* **2005**, *122*, 204108. Luis, J. M.; Torrent-Sucarrat, M.; Christiansen, O.; Kirtman, B. *J. Chem. Phys.* **2007**, *127*, Zalesny, R.; Papadopoulos, M. G.; Bartkowiak, W.; Kaczmarek, A. *J. Chem. Phys.* **2008**, *129*, 134310. Chou, C. C.; Jin, B. Y. *Chem. Phys.* **2009**, *362*, 71. Chou, C. C.; Jin, B. Y. *Theor. Chem. Acc.* **2009**, *122*, 313. Luis, J. M.; Reis, H.; Papadopoulos, M.; Kirtman, B. *J. Chem. Phys.* **2009**, *131*, 034116. Zalesny, R.; Wojcik, G.; Mossakowska, I.; Bartkowiak, W.; Avramopoulos, A.; Papadopoulos, M. G. *THEOCHEM* **2009**, *907*, 46.
- (24) Torrent-Sucarrat, M.; Solà, M.; Duran, M.; Luis, J. M.; Kirtman, B. *J. Chem. Phys.* **2002**, *116*, 5363.
- (25) Torrent-Sucarrat, M.; Solà, M.; Duran, M.; Luis, J. M.; Kirtman, B. *J. Chem. Phys.* **2003**, *118*, 711.
- (26) Torrent-Sucarrat, M.; Solà, M.; Duran, M.; Luis, J. M.; Kirtman, B. *J. Chem. Phys.* **2004**, *120*, 6346.
- (27) Loboda, O.; Zalesny, R.; Avramopoulos, A.; Luis, J. M.; Kirtman, B.; Tagmatarchis, N.; Reis, H.; Papadopoulos, M. G. *J. Phys. Chem. A* **2009**, *113*, 1159.
- (28) Champagne, B.; Perpete, E. A.; van Gisbergen, S. J. A.; Baerends, E. J.; Snijders, J. G.; Soubra-Ghaoui, C.; Robins, K. A.; Kirtman, B. *J. Chem. Phys.* **1998**, *109*, 10489. van Gisbergen, S. J. A.; Schipper, P. R. T.; Gritsenko, O. V.; Baerends, E. J.; Snijders, J. G.;

- Champagne, B.; Kirtman, B. *Phys. Rev. Lett.* **1999**, *83*, 694. Champagne, B.; Perpete, E. A.; Jacquemin, D.; van Gisbergen, S. J. A.; Baerends, E. J.; Soubra-Ghaoui, C.; Robins, K. A.; Kirtman, B. *J. Phys. Chem. A* **2000**, *104*, 4755.
- (29) Suponitsky, K. Y.; Liao, Y.; Masunov, A. E. *J. Phys. Chem. A* **2009**, *113*, 10994.
- (30) Leininger, T.; Stoll, H.; Werner, H. J.; Savin, A. *Chem. Phys. Lett.* **1997**, *275*, 151. Iikura, H.; Tsuneda, T.; Yanai, T.; Hirao, K. *J. Chem. Phys.* **2001**, *115*, 3540. Toulouse, J.; Colonna, F.; Savin, A. *Phys. Rev. A* **2004**, *70*. Tawada, Y.; Tsuneda, T.; Yanagisawa, S.; Yanai, T.; Hirao, K. *J. Chem. Phys.* **2004**, *120*, 8425. Chai, J. D.; Head-Gordon, M. *J. Chem. Phys.* **2008**, *128*, 084106. Chai, J. D.; Head-Gordon, M. *Phys. Chem. Chem. Phys.* **2008**, *10*, 6615.
- (31) Kamiya, M.; Sekino, H.; Tsuneda, T.; Hirao, K. *J. Chem. Phys.* **2005**, *122*. Peach, M. J. G.; Cohen, A. J.; Tozer, D. J. *Phys. Chem. Chem. Phys.* **2006**, *8*, 4543. Borini, S.; Limacher, P. A.; Luthi, H. P. *J. Chem. Phys.* **2009**, *131*, 123105. Bonness, S.; Fukui, H.; Yoneda, K.; Kishi, R.; Champagne, B.; Botek, E.; Nakano, M. *Chem. Phys. Lett.* **2010**, *493*, 195. Kishi, R.; Bonness, S.; Yoneda, K.; Takahashi, H.; Nakano, M.; Botek, E.; Champagne, B.; Kubo, T.; Kamada, K.; Ohta, K.; Tsuneda, T. *J. Chem. Phys.* **2010**, *132*, 094107. de Wergifosse, M.; Champagne, B. *J. Chem. Phys.* **2011**, *134*, 074113.
- (32) Jacquemin, D.; Perpete, E. A.; Medved, M.; Scalmani, G.; Frisch, M. J.; Kobayashi, R.; Adamo, C. *J. Chem. Phys.* **2007**, *126*, 191108. Jacquemin, D.; Perpete, E. A.; Scalmani, G.; Frisch, M. J.; Kobayashi, R.; Adamo, C. *J. Chem. Phys.* **2007**, *126*, 144105. Lu, S. I. *J. Comput. Chem.* **2010**, *32*, 730.
- (33) Limacher, P. A.; Mikkelsen, K. V.; Luthi, H. P. *J. Chem. Phys.* **2009**, *130*, 194114.
- (34) Rivera-Fuentes, P.; Aonso-Gomez, J. L.; Petrovic, A. G.; Seiler, P.; Santoro, F.; Harada, N.; Berova, N.; Rzepa, H. S.; Diederich, F. *Chem.—Eur. J.* **2011**, *16*, 9796. Cherblanc, F.; Lo, Y.; De Gussem, E.; Alcazar-Fuoli, L.; Bignell, E.; He, Y.; Chapman-Rothe, N.; Bultinck, P.; Herrebout, W. A.; Brown, R.; Rzepa, H. S.; Fuchter, M. J. *Chem. Eur. J.* DOI: 10.1002/chem.201101129.
- (35) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648.
- (36) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 1372.
- (37) Boese, A. D.; Martin, J. M. L. *J. Chem. Phys.* **2004**, *121*, 3405.
- (38) Zhao, Y.; Schultz, N. E.; Truhlar, D. G. *J. Chem. Theory Comput.* **2006**, *2*, 364.
- (39) Yanai, T.; Tew, D. P.; Handy, N. C. *Chem. Phys. Lett.* **2004**, *393*, 51.
- (40) Møller, C.; Plesset, M. S. *Phys. Rev.* **1934**, *46*, 618.
- (41) Hehre, W. J.; Ditchfield, R.; Pople, J. A. *J. Chem. Phys.* **1972**, *56*, 2257. Hehre, W. J.; Radom, L.; Schleyer, P. v. R.; Pople, J. A. *Ab Initio Molecular Orbital Theory*; Wiley: New York, 1986.
- (42) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, N. J.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, Ö.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. *Gaussian 09*, Revision A.1; Gaussian, Inc.: Wallingford, CT, 2009.
- (43) The Hückel and Möbius conformations of bianthraquinodimethane modified [16] annulene show very small dipole moments, and we expect that the solvent will not play a crucial role in the values of NLOP.
- (44) Davis, P. J.; Rabinowitz, P. In *Numerical Integration*; Blasdell: London, 1967; p 166. Champagne, B.; Kirtman, N. In *Handbook of Advanced Electronic and Photonic Materials*; Nalwas, H. S., Ed.; Academic: San Diego, 2001; Vol. 9, p 63.
- (45) Bishop, D. M.; Kirtman, B. *J. Chem. Phys.* **1991**, *95*, 2646. Bishop, D. M.; Kirtman, B. *J. Chem. Phys.* **1992**, *97*, 5255. Bishop, D. M.; Luis, J. M.; Kirtman, B. *J. Chem. Phys.* **1998**, *108*, 10013.
- (46) Christiansen, O. *J. Chem. Phys.* **2005**, *122*, 194105. Christiansen, O.; Kongsted, J.; Paterson, M. J.; Luis, J. M. *J. Chem. Phys.* **2006**, *125*, 214309.
- (47) Bishop, D. M.; Hasan, M.; Kirtman, B. *J. Chem. Phys.* **1995**, *103*, 4157. Kirtman, B.; Luis, J. M.; Bishop, D. M. *J. Chem. Phys.* **1998**, *108*, 10008.
- (48) Luis, J. M.; Marti, J.; Duran, M.; Andres, J. L.; Kirtman, B. *J. Chem. Phys.* **1998**, *108*, 4123.
- (49) Bishop, D. M.; Dalskov, E. K. *J. Chem. Phys.* **1996**, *104*, 1004. Quinet, O.; Champagne, B. *J. Chem. Phys.* **1998**, *109*, 10594.
- (50) Luis, J. M.; Duran, M.; Kirtman, B. *J. Chem. Phys.* **2001**, *115*, 4473.
- (51) Luis, J. M.; Duran, M.; Andres, J. L.; Champagne, B.; Kirtman, B. *J. Chem. Phys.* **1999**, *111*, 875. Luis, J. M.; Champagne, B.; Kirtman, B. *Int. J. Quantum Chem.* **2000**, *80*, 471.
- (52) Luis, J. M.; Duran, M.; Champagne, B.; Kirtman, B. *J. Chem. Phys.* **2000**, *113*, 5203.
- (53) Kirtman, B.; Bonness, S.; Ramirez-Solis, A.; Champagne, B.; Matsumoto, H.; Sekino, H. *J. Chem. Phys.* **2008**, *128*, 114108.

Dispersion Interactions with Density-Functional Theory: Benchmarking Semiempirical and Interatomic Pairwise Corrected Density Functionals

Noa Marom,^{†,‡} Alexandre Tkatchenko,[§] Mariana Rossi,[§] Vivekanand V. Gobre,[§] Oded Hod,^{||} Matthias Scheffler,[§] and Leeor Kronik^{*,†}

[†]Department of Materials and Interfaces, Weizmann Institute of Science, Rehovoth 76100, Israel

[‡]Center for Computational Materials, Institute for Computational Engineering and Sciences, University of Texas at Austin, Austin, Texas 78712, United States

[§]Fritz-Haber-Institut der Max-Planck-Gesellschaft, Faradayweg 4-6, 14195 Berlin, Germany

^{||}School of Chemistry, The Sackler Faculty of Exact Sciences, Tel Aviv University, Tel Aviv 69978, Israel

ABSTRACT: We present a comparative assessment of the accuracy of two different approaches for evaluating dispersion interactions: interatomic pairwise corrections and semiempirical meta-generalized-gradient-approximation (meta-GGA)-based functionals. This is achieved by employing conventional (semi)local and (screened-)hybrid functionals, as well as semiempirical hybrid and nonhybrid meta-GGA functionals of the M06 family, with and without interatomic pairwise Tkatchenko–Scheffler corrections. All of those are tested against the benchmark S22 set of weakly bound systems, a representative larger molecular complex (dimer of NiPc molecules), and a representative dispersively bound solid (hexagonal boron nitride). For the S22 database, we also compare our results with those obtained from the pairwise correction of Grimme (DFT-D3) and nonlocal Langreth–Lundqvist functionals (vdW-DF1 and vdW-DF2). We find that the semiempirical kinetic-energy-density dependence introduced in the M06 functionals mimics some of the nonlocal correlation needed to describe dispersion. However, long-range contributions are still missing. Pair-wise interatomic corrections, applied to conventional semilocal or hybrid functionals, or to M06 functionals, provide for a satisfactory level of accuracy irrespectively of the underlying functional. Specifically, screened-hybrid functionals such as the Heyd–Scuseria–Ernzerhof (HSE) approach reduce self-interaction errors in systems possessing both localized and delocalized orbitals and can be applied to both finite and extended systems. Therefore, they serve as a useful underlying functional for dispersion corrections.

1. INTRODUCTION

Dispersive interactions are essential to understanding many important phenomena in chemistry, biology, and materials science. Such interactions possess a significant attractive component due to instantaneous dipoles and higher-order multipoles. They typically dominate in regions where there is little or no overlap of electron densities, i.e., at medium to long range, as compared to the short-range covalent and ionic bonds.¹ There is presently great interest in treating dispersive interactions correctly within density-functional theory (DFT), which has become the method of choice for electronic-structure calculations across an unusually wide variety of fields, from organic chemistry to solid-state physics.^{2,3} In principle, DFT is an exact theory, and the exact functional must also include an accurate treatment of dispersive interactions. In practice, from a DFT perspective, dispersive interactions are a long-range correlation phenomenon that is very difficult to account for accurately. Indeed, determining appropriate long-range correlation expressions without up-setting the delicate balance between exchange and correlation contributions is a highly nontrivial task,^{4–15} often associated with a considerable computational cost. Therefore, distinctly different strategies for *circumventing* the explicit evaluation of long-range correlation have emerged.

One popular strategy for describing both strong and dispersive chemical interactions is to augment conventional functionals with pairwise addition of C_6/R^6 correction terms to the inter-nuclear energy expression.^{16–25} These terms are smoothly cut off in the short range, where they are not relevant, but explicitly enforce the desired long-range asymptotic behavior. A different strategy is to use semilocal or hybrid functionals that contain a large number of free parameters in the functional form. These parameters are semiempirically fit using diverse data sets that include data pertinent not only to thermochemistry but also to noncovalent interactions, kinetics, and more. In this way, many deficiencies of traditional semilocal and hybrid functionals, including the treatment of dispersion, can be minimized. This philosophy is best exemplified by the M06 suite of functionals^{26,27}—a “family” of four (possibly) hybrid meta-GGA functionals (i.e., functionals that also depend explicitly on the kinetic-energy density²⁸). The M06 functionals differ in the fraction of exact exchange used—zero (M06-L, a semilocal functional), a fraction similar to that of standard hybrid functionals (M06), a fraction twice as large (M06–2X), and one (M06-HF).

Received: August 11, 2011

Published: October 27, 2011

Both strategies have been put to extensive use in the past few years for accurate structural predictions of various properties of dispersively bound systems (see, e.g., refs 29–50). Given the significant differences between them, it is important to assess the level of accuracy that can be expected from both. Here, we do so by comparing systematically the performance of both methods on prototypical small, large, and extended dispersively bound systems. In particular, we examine selected conventional semi-local and hybrid functionals, as well as representative semi-empirical hybrid and nonhybrid meta-generalized-gradient-approximation (meta-GGA) functionals. All of these functionals are used both with and without C_6/R^6 correction terms, allowing a direct identification of the important role played by long-range interactions.

2. METHODOLOGY

We have selected the following representatives of important classes of functionals for our comparison. These include (i) the earliest practical approximate density functional—the local-density approximation (LDA);^{51–53} (ii) the nonempirical generalized gradient approximation (GGA) functional of Perdew, Burke, and Ernzerhof (PBE);^{54,55} (iii) two conventional hybrid functionals—the nonempirical PBE0 functional, where 25% of exact exchange is admixed with the semilocal exchange,^{56–58} and the semiempirical B3LYP functional,^{59,60} where 20% of exact exchange are admixed, together with further semiempirical modifications of semilocal exchange and correlation; (iv) the screened hybrid functional of Heyd, Scuseria, and Ernzerhof (HSE),^{61,62} where only short-range exact-exchange is admixed, with the aid of an empirical range-separation parameter, such that it approaches PBE0 in the short-range and PBE in the long-range; (v) the two most commonly used representatives of the M06 semiempirical functionals—the meta-GGA M06-L functional and the hybrid meta-GGA M06 functional;^{26,27} and (vi) the nonempirical meta-GGA functional of Tao, Perdew, Staroverov, and Scuseria (TPSS).⁶³ LDA, PBE, B3LYP/PBE0, and HSE were chosen because they are, respectively, local, semilocal, conventional hybrid, and short-range hybrid functionals whose construction did not involve explicit consideration of dispersive interactions. M06 and M06-L were chosen to represent, respectively, hybrid and nonhybrid flavors of semiempirical functionals constructed with dispersive interactions in mind. Finally, the nonempirical TPSS meta-GGA functional was additionally employed in some comparisons in order to assess the importance of empiricism in the M06 meta-GGA-based constructs.

All functionals except LDA were considered both with and without pairwise corrections. (We exclude the LDA dispersion correction as LDA exhibits short-range overbinding,⁶⁴ and Karton et al. have shown that if one insists on fitting C_6 coefficients against the S22 set anyway, unphysical negative values are obtained.⁶⁵) Here, we used the pairwise correction scheme suggested by Tkatchenko and Scheffler (TS-vdW).²⁵ As in other such schemes,^{16–24} in the TS-vdW approach, the dispersion correction energy, E_{disp} added to the internuclear energy term, is given by

$$E_{\text{disp}} = - \sum_{j > i} f_{\text{damp}}(R_{ij}, R_{ij}^0) C_{6ij} R_{ij}^{-6} \quad (1)$$

where C_{6ij} is the dispersion coefficient for the ij pair of atoms, R_{ij} is the interatomic distance, R_{ij}^0 is the sum of equilibrium vdW radii for the pair, and f_{damp} is a damping function. The latter is

chosen in the form of a Fermi–Dirac function:

$$f_{\text{damp}}(R_{ij}, R_{ij}^0) = \left[1 + \exp \left(-d \left(\frac{R_{ij}}{s_R R_{ij}^0} - 1 \right) \right) \right]^{-1} \quad (2)$$

where d determines the “steepness” of the damping and s_R determines the range of damping—the larger s_R is, the larger the range of interaction for which dispersion is already well-handled by the underlying exchange–correlation functional. Distinctly from previous pairwise corrections schemes, in the TS-vdW scheme, the parameters $C_{6ij}[n(r)]$ and $R_{ij}^0[n(r)]$ are functionals of the electron density $n(r)$, as they take into account the relative volume of each atom inside the system, based on Hirshfeld⁶⁶ partitioning. Importantly, the range parameter s_R is the only one that needs to be determined empirically. This is achieved by fitting s_R for each underlying functional, once and for all, to the S22 data set of Jurečka et al.⁶⁷ This set contains binding energies of 22 different weakly bound systems, calculated using the coupled cluster method with single, double, and triple excitations, where triple excitations are treated perturbatively [CCSD(T)], with a numerical accuracy close to the basis set limit. Here, the basis-set extrapolated CCSD(T) values, reported recently by Sherrill et al.,⁶⁸ were used.

The TS-vdW correction has been implemented in FHI-aims,^{69,70} an all-electron electronic structure code which employs efficient numerical atom-centered orbitals (NAO) as a basis set. In this work, the tier2 NAO basis set has been employed throughout. This basis set yields results that are similar in accuracy to those of the aug-cc-pVQZ Gaussian basis set for the S22 database²⁵ and has been explicitly tested for convergence by selected comparisons with computations using the higher level, tier3 NAO basis set.

The M06 suite of functionals was implemented into the FHI-aims code in a postprocessing fashion. The exchange–correlation energy with the M06 functional of choice is calculated using all quantities (density, density gradient, and kinetic energy density) derived from the self-consistent PBE orbitals. This approach was validated against the self-consistent M06 implementation, found in the Gaussian code (G09),⁷¹ for the S22 set. Binding energies computed with the two approaches generally exhibited differences that were smaller than 4 meV (the largest difference is of 8 meV for a total binding energy of ~ 780 meV). Correction for basis-set superposition errors (BSSE) resulted in differences on the same order of magnitude. The typical total energy deviation between G09 and FHI-aims is 0.01 Hartree (0.005%). Additionally, G09 was used with triple- and quadruple- ζ quality basis sets for geometry optimizations of the *h*-BN sheet with all functionals and of the NiPc monomer with the hybrid functionals and M06L.

3. RESULTS AND DISCUSSION

An important figure of merit for the performance of the different functionals tested here is their performance on the benchmark S22 set. As a first step, the range parameter, s_R (see eq 2), was determined using the S22 set based on the recently reported basis-set extrapolated coupled-cluster data of Sherrill et al.,⁶⁸ which exhibit a mean absolute relative error of $\sim 2\%$ with respect to the Jurečka et al. data.⁶⁷ The s_R values obtained from this procedure are compiled in Table 1. The mean absolute errors obtained from all functionals, with and without pairwise corrections, are shown in Figure 1a.

Table 1. Range Parameters of the TS-vdW Correction for Different Functionals and the Mean Absolute Errors in meV with and without the TS-vdW Correction with Respect to CCSD(T) Reference Values⁶⁸ for the Binding Energies of the S22 Data Set (Also Shown Graphically in Figure 1)

functional	s_R	MAE without the TS-vdW correction				MAE with the TS-vdW correction			
		H-bond	vdW	mixed	total	H-bond	vdW	mixed	total
LDA	N/A	219	16	42	89	N/A	N/A	N/A	N/A
PBE	0.94	53	208	87	118	16	14	7	12
PBEh	0.96	41	191	78	107	22	10	7	13
HSE	0.96	36	185	76	103	31	12	7	17
B3LYP	0.84	89	266	125	165	11	12	6	10
TPSS	0.86	87	250	118	156	14	7	7	9
M06L	1.27	22	27	38	29	15	19	13	16
M06	1.16	52	70	52	59	27	14	14	18

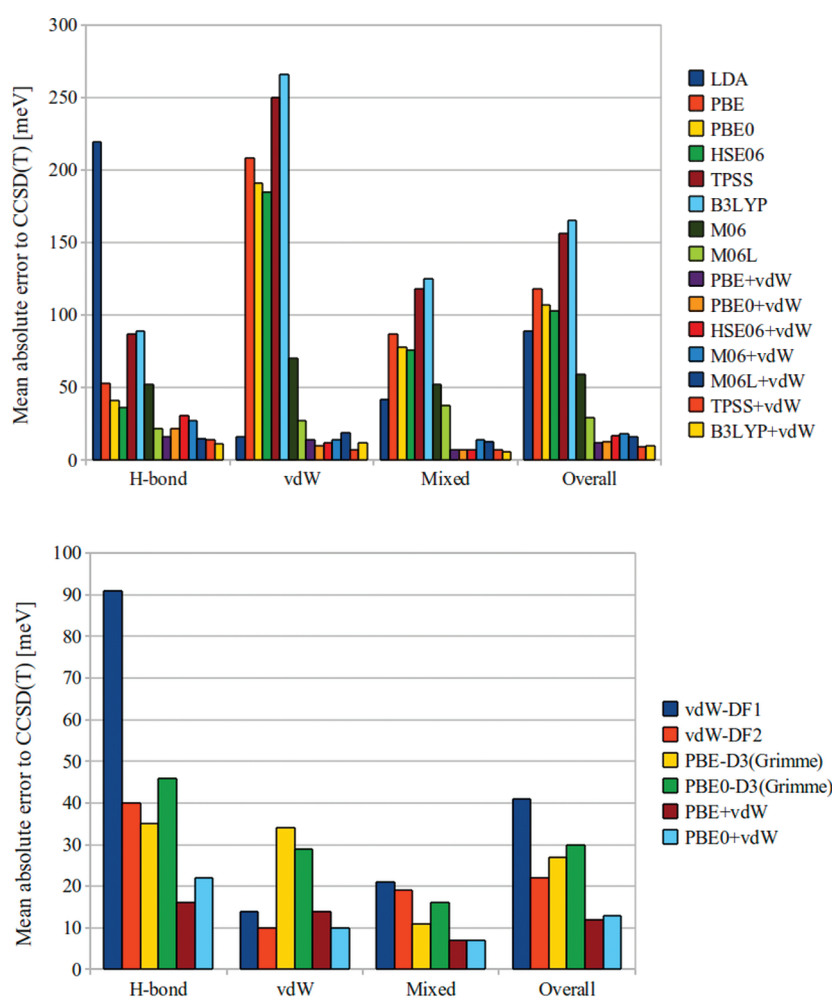


Figure 1. (a) Mean absolute errors of different functionals with and without the TS-vdW correction with respect to CCSD(T) reference values⁶⁸ for the binding energies of the S22 data set. For each of the TS-vdW corrected results, the calculations employed the empirical parameter s_R reported in Table 1, which was optimized on the basis of the S22 set. All of the “+vdW” results were obtained with the per functional optimal s_R parameter reported in Table 1. (b) Mean absolute errors of two of the TS-vdW corrected functionals, compared to the mean absolute errors of the same two functionals corrected by the D3 Grimme pairwise scheme and to the mean absolute errors of both variants of the Langreth–Lundqvist “DF-vdW” approach.

Several conclusions can be immediately drawn from Figure 1a: First, LDA does remarkably well for the dispersively bound systems. However, as we demonstrate further below, this simply reflects its short-range overbinding nature,⁶⁴ which provides for

error cancellation with the missing electron correlation. At the same time, LDA does very poorly for the hydrogen-bonded systems in the S22 set, and so its overall performance is unsatisfactory. For the three PBE-based functionals—PBE, PBE0

(conventional hybrid), and HSE (short-range hybrid)—the opposite picture emerges. All three functionals do quite poorly for the dispersively bound systems. However, they do significantly better than LDA (though still do not offer chemical accuracy) for the hydrogen-bonded systems, due to a better description of the polarization. Furthermore, the transition from PBE to PBE0 or to HSE offers some improvement of accuracy, but this improvement is too small to be of any practical significance. As for the semiempirical hybrid, B3LYP, it too does better than LDA for hydrogen-bonded systems and worse than LDA for dispersion. However, its performance in either category is markedly worse than that of the nonempirical PBE0 functional. Likely, this is a consequence of the absence of weakly bound systems from the training set used to optimize the B3LYP functional.

Clearly, the empirically parametrized meta-GGA-based functionals offer a significant improvement. M06 markedly improves dispersion energies while performing similarly to the PBE-based functionals for hydrogen-bonding. M06-L offers a further and considerable reduction of mean errors for both kinds of weak bonding, resulting in an overall better accuracy than all functionals surveyed so far. Thus, the M06 functionals do indeed capture more of the correlation relevant for weak, nonlocal interactions than that captured by the standard semilocal and hybrid functionals. This agrees well with previous reports of their success in describing weakly bound systems.^{43–50} Furthermore, comparison of the M06-L results with those of the nonempirical TPSS meta-GGA demonstrate unequivocally that the M06 approach owes its success to the empirical construction rather than to the meta-GGA construct. In fact, the TPSS results are markedly worse than those of M06-L and are surprisingly close to those of B3LYP, despite the considerable differences between these two functionals.

Despite this significant advance, Figure 1a shows that there is room for even further improvement. In fact, all seven pairwise-corrected functionals offer a further ~50% reduction in mean absolute overall error over M06-L, mostly due to improved accuracy for the dispersive and mixed binding complexes. This shows unequivocally that while the use of empirical meta-GGA expressions does mimic successfully some of the missing long-range correlation (likely by incorporation of “middle-ranged” correlation), its functional form is still limited and notably is not asymptotically correct. Therefore, it still benefits from pairwise interatomic corrections.

The same trends are also reflected in the range parameter s_R of Table 1. The smaller this parameter is, the smaller the damping of pairwise-correction is (see eq 2). In other words, a smaller value implies a “deeper” pairwise correction. Accordingly, B3LYP and TPSS, which do most poorly for dispersion, need the largest pairwise correction and indeed have the smallest s_R . Next, PBE, PBE0, and HSE have a larger, and very similar, s_R . This is because, as discussed above, their handling of dispersion is similar. Finally, M06 exhibits a markedly larger s_R , and M06-L, which does best for dispersion, has the largest s_R . This reaffirms the above “ranking” of functional performance and provides a complementary perspective on its reasons. But perhaps more importantly, it shows that the single semiempirical parameter s_R is sufficient for making sure that dispersive corrections are added without “double counting” of electron correlation. Consequently, a similar degree of accuracy is achieved for any reasonable underlying exchange-correlation functional. Importantly, we note that the TS-vdW correction has a negligible effect for covalent

bonds—on average, it changes the atomization energies of small organic molecules in the G2 data set by only ~0.1 kcal/mol.

At this point, one may wonder whether the significant improvement in accuracy for the S22 set, obtained by applying TS-vdW corrections (especially with respect to the M06 family of functionals), does not merely reflect the fact that the TS-vdW range parameter, s_R , was fit against the same data set. To exclude that, we have performed additional calculations with PBE and M06, with and without TS-vdW corrections, for the larger and more diverse, recently suggested S66 set.⁷² We find that without TS-vdW corrections, the mean average error (in meV) is 96 with PBE and a significantly smaller 43 with M06. After application of the TS-vdW correction, the errors are further reduced for both functionals and become very close to each other: 19 and 21 for the PBE- and M06-based dispersion-corrected calculations, respectively. These numbers follow the same trends observed in Figure 1 and are quantitatively close to those reported in Table 1 for the smaller S22 set. This firmly establishes that the above conclusions remain valid outside the original training set used for determining the TS-vdW range parameter.

Before considering systems beyond standard benchmark sets, it is instructive to compare these benchmark data to results obtained using two other highly popular methods for considering dispersive interactions within DFT. The first is the most recent pairwise correction suggested by Grimme and co-workers (DFT-D3).²¹ The second method we compare our results to is the Langreth–Lundqvist “vdW-DF” approach, in both its original (vdW-DF1)⁹ and recently revised (vdW-DF2)¹⁰ forms. This approach differs from those discussed so far by explicitly introducing a nonlocal correlation expression, which is combined with local correlation to get the total correlation expression. In Figure 1b, we compare the mean absolute errors for the S22 set, obtained from the PBE and PBE0 functionals with both DFT-D3 and TS-vdW pairwise corrections, as well as from the vdW-DF1 and vdW-DF2 approaches.^{12,21} For this set, the TS-vdW approach used with PBE and PBE0 functionals is, on average, twice as accurate as the DFT-D3, vdW-DF1, and vdW-DF2 approaches. Several comments are, however, in order. First, in the DFT-D3 approach, empirical parameters were obtained on the basis of a rich training set including, but not limited to, the S22 set. Possibly, optimization based on the S22 set alone may have led to somewhat improved performance for this set (although, as shown above, the accuracy of the TS-vdW approach extends beyond the S22 set). Furthermore, the DFT-D3 approach tends to perform better with Becke-type exchange (as in, e.g., the BLYP or B3LYP functionals).²¹ Here, however, we prefer to focus on PBE-related functionals because we are interested in minimal empiricism. Third, the introduction of a damping function in a vdW-DF-like approach¹¹ can lead to accuracy that is substantially better than both vdW-DF1 and vdW-DF2 methods for the S22 set,⁷³ at the expense of adding more empiricism.^{74–76} Despite these caveats, the comparison of Figure 1b is important, because (1) it involves some of the most commonly used alternate methods for incorporating dispersion interactions, in their original form, and (2) it focuses on functionals that have been applied with similar accuracy in both molecular and solid-state systems.

We now return to the comparison of TS-vdW to the M06 family. In order to examine whether the observations made above for the S22 set, against which the pairwise corrections were explicitly fit, hold for systems outside the set, we consider two prototypical dispersively bound systems: a molecular complex,

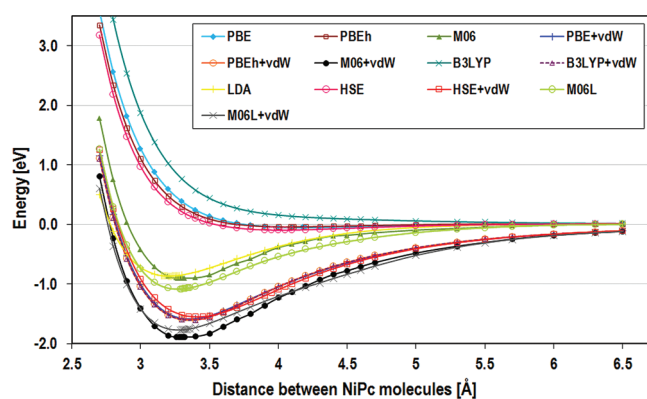


Figure 2. Binding energy curves of the NiPc dimer obtained with different functionals with and without the TS-vdW correction.

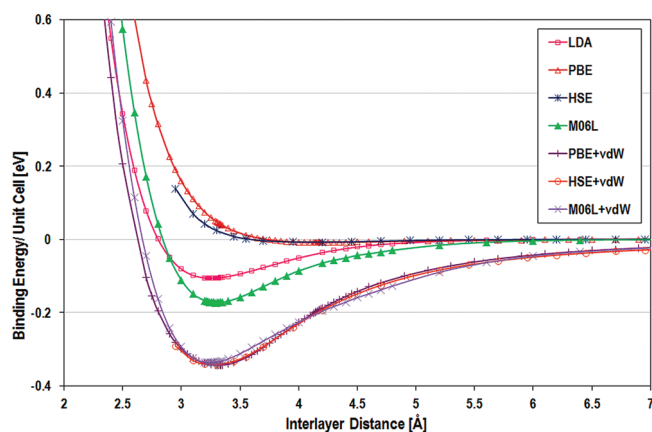


Figure 3. Binding energy curves of bulk hBN with respect to a single hBN sheet, obtained with different functionals with and without the TS-vdW correction.

the Ni–phthalocyanine (NiPc) dimer, and a layered solid, hexagonal boron nitride (hBN). These systems were chosen because both the TS-vdW method and the M06 family^{29,30,46} were previously shown to yield highly satisfactory predictions for the geometry of these systems, making a direct comparison of energetics intriguing.

Binding energy curves obtained with all of the above functionals for the NiPc dimer, as a function of the intermonomer distance, are shown in Figure 2. The monomer geometry was reoptimized with each functional. For hBN, the hybrid functionals B3LYP, PBE0, and M06 (with or without pairwise corrections) were not considered due to the computational difficulties associated with employing them in a periodic system. Binding energy curves obtained with all remaining functionals, as a function of the interlayer distance (without further intralayer relaxation), are shown in Figure 3. The geometry of the hBN sheet was reoptimized with each functional. The equilibrium distances obtained from these binding curves are summarized in Table 2.

The different binding curves of Figures 2 and 3 can be roughly divided into three distinct groups. In the first group are the standard functionals PBE, PBE0, HSE, and B3LYP, which predict no binding at all (B3LYP) or very little binding with unrealistically small binding energy and unrealistically large interunit distance. In the second group, one finds M06 and M06-L, and in the

Table 2. Equilibrium Distances of the NiPc Dimer and Bulk hBN Obtained with Different Functionals with and without the TS-vdW Correction, Compared to Experimental Results^a

functional	equilibrium distance [Å]	
	NiPc	hBN
PBE	4.2	4.17
PBEh	4.1	N/A
HSE	4	4.2
M06	3.30	N/A
M06L	3.29	3.30
LDA	3.21	3.25
PBE+vdW	3.4	3.33
PBEh+vdW	3.4	N/A
HSE+vdW	3.4	3.3
M06+vdW	3.29	N/A
M06L+vdW	3.29	3.25
experiment	3.24 ⁹⁴	3.33 ⁹⁵

^aNote that the values given here for the NiPc dimer were not obtained from full geometry optimization of the dimer but rather from varying the intermolecular distance with the monomer geometry fixed. Complete relaxation improves agreement with experiment³⁰al results.

third group, one finds all pairwise-corrected functionals. For both of these groups, Table 2 shows that satisfying agreement with experimental results is found for the geometry, in agreement with previous studies. These two groups differ by roughly a factor of 2, however, in their estimate of the binding energy. Interestingly, the LDA results are much closer to the M06(-L) ones than to the TS-vdW ones. However, the underestimated equilibrium distances (Table 2) and, for hBN, the clear further underestimate of the binding energy show conclusively that LDA is not a suitable functional for investigations of dispersively bound systems, despite its reasonable performance for the dispersively bound systems in the S22 set.

In the absence of experimental binding energy data, which group of results should we trust? For the S22 set, we found that the M06(-L) predictions are always an underestimate of the coupled cluster binding energy. In contrast, pairwise-corrected calculations sometimes overestimate and sometimes underestimate the coupled cluster result, with a much smaller mean absolute error. In light of this, as well as the fact that the missing long-range contributions must gain in significance as the system size grows, we expect that the M06 curves systematically underestimate the true binding energies. Nevertheless, one may equally well expect that the TS-vdW approach can overestimate the true binding energy, especially for the solid state system. This is because the TS-vdW approach used here lacks a proper description for the screening of the pairwise interaction by the dielectric medium.^{77–80} Nevertheless, owing to its overall better performance, we expect that the true binding energy is closer to the TS-vdW results than to the M06 result. Thus, we again conclude that even if one chooses to use the M06(-L) approach, further pairwise corrections are still desired. That said, if TS-vdW corrections are utilized, one may reach the same level of accuracy, by employing the TS-vdW corrections in conjunction with a conventional semilocal or hybrid functional.

At this point, some notes on the “ladder” of the incorporation of long-range correlation are in order. First, consider that for a uniform electron gas the LDA functional, as well as the PBE and

PBE0 functionals, which reduce to LDA in this limit, are *exact*. Therefore, they must describe all relevant physical phenomena, including dispersion.¹⁹ However, for a nonuniform system, the absence of a nonlocal correlation term inherently does not allow for the instantaneous motion of electrons in one part of the system to be correlated with that in another part. Hence, the ability to describe dispersive interactions by (semi)-local functionals is lost in the transition from a uniform system to a nonuniform one. Conventional hybrid functionals are not expected to help in this respect, because it is long-range correlation, rather than exchange, that governs these weak interactions. This immediately explains why PBE, PBE0, and HSE, all of which are based on PBE correlation, are not fundamentally different from each other for describing dispersive interactions in general and for the S22 set in particular. B3LYP fares even worse owing to parametrization not being oriented toward weakly bound systems.

Meta-GGA functionals are often referred to as semilocal, but here it is important to notice that meta-GGA functionals may contain a kinetic energy density term is semilocally dependent on the Kohn–Sham orbitals rather than on the density.^{28,81} Generally, a functional that is semilocal in the orbitals need not be semilocal in the density, as the link between orbitals and density is established by solving the Kohn–Sham equations for the orbital-dependent functional. It is this link which allows the introduction of density-dependent nonlocality, and in this sense the meta-GGA approach is not semilocal.²⁸ The success of M06 and M06-L can then be viewed as the “maximization” of the extent of nonlocal correlation that can be established in this way, via empirical construction (as demonstrated above by its comparison to the nonempirical TPSS meta-GGA for the S22 set). However, nonlocal correlation expressed through the kinetic energy still requires some orbital overlap,⁴⁶ and ergo it too decays too rapidly. This explains why even though the M06 family was explicitly constructed to consider dispersion, it too still lacks true long-range correlation and still gains from interatomic correlation corrections, as shown above and in refs 30 and 65.

Finally, we have previously shown that a significant advantage of the TS-vdW scheme is that the difficult problem of obtaining a description of both geometry and electronic structure can be generally overcome by decoupling the two issues.³⁰ One can choose a functional that is appropriate for the electronic structure but does not necessarily include a good description of dispersive interactions and augment it with first-principles corrections for the leading terms of the dispersion interaction using the TS-vdW approach. Figures 1–3 show that the degree to which the binding energy is well-described is quite weakly dependent on the functional underlying the interatomic corrections. This is important, because it is by now well-established that the electronic structure obtained from hybrid functionals (including both HSE and M06) is superior to that obtained from semilocal ones (including M06-L) whenever localized and delocalized orbitals need to be described on the same footing (see, e.g., refs 30, 82–92). The present results indicate that, especially for extended systems, TS-vdW-corrected HSE calculations then emerge as the current method of choice for obtaining a good description of both electronic structure and dispersion interactions. This is because, on the one hand, the short-range exchange is sufficient for mitigating the self-interaction errors associated with orbital localization and, on the other hand, the elimination of the long-range exchange keeps the computational cost reasonably low. Finally, we note that further progress is possible. As mentioned above, the herewith

discussed dispersion-correction approaches, including TS-vdW, lack a full microscopic description of dielectric screening and nonadditive many-body vdW energy contributions for large molecules and solids. The description of these two effects is the subject of ongoing research.⁹³

4. CONCLUSIONS

In conclusion, we have presented a comprehensive evaluation of the performance of representative (semi)local, (screened-) hybrid, and semiempirical meta-GGA functionals, with and without interatomic pairwise TS-vdW corrections, for the dispersively bound S22 set, a dimer of NiPc molecules, and the layered solid hexagonal boron nitride. Clear and distinct trends are identified in all cases. LDA can mimic van der Waals attraction, though its equilibrium predictions can be “right for the wrong reason”, a fact easily exposed when considering the full binding energy curve. Conventional semilocal functionals describe dispersive attraction very poorly. Conventional or short-range hybrids based on these semilocal functionals perform almost exactly the same way, as they do not improve the treatment of long-range correlation. More nonlocal correlation is introduced via semiempirical kinetic-energy-density dependence, as in the M06 functionals, but some long-range contributions are still missing. Pair-wise interatomic corrections improve binding energies throughout, even when applied to M06 calculations, but the same level of accuracy (which, for the S22 set exceeds both the DFT-D3 and the vdW-DF methods) can be reached by applying these corrections to conventional semilocal or hybrid functionals. The correlation trends across the different functionals are fully mirrored by the range at which the dispersive corrections are dampened. Finally, pairwise-corrected functionals generally allow for separate optimization of the electronic structure problem and the dispersive interaction problem. Specifically, vdW-corrected HSE calculations emerge as a computationally tractable means for assessing dispersively bound solids possessing both localized and delocalized orbitals.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: leeor.kronik@weizmann.ac.il.

■ ACKNOWLEDGMENT

Work at the Weizmann Institute was supported by the Israel Science Foundation, the Gerhard Schmidt Minerva Center for Supra-Molecular Architecture, the Lise Meitner Center for Computational Chemistry, and the historical generosity of the Perlman family. Work at Tel-Aviv University was supported by the Israel Science Foundation under grant no. 1313/08, the European Community’s Seventh Framework Programme FP7/2007-2013 under grant agreement no. 249225, the Center for Nanoscience and Nanotechnology at Tel Aviv University, and the Lise-Meitner Center for Computational Chemistry. A.T. acknowledges financial support from the Alexander von Humboldt (AvH) foundation.

■ REFERENCES

- (1) Parsegian, V. A. *Van der Waals Forces: A Handbook for Biologists, Chemists, Engineers, and Physicists*; Cambridge University Press: Cambridge, U. K., 2006.

- (2) Koch, W.; Holthausen, M. C. *A Chemist's Guide to Density Functional Theory*; Wiley-VCH: Weinheim, Germany, 2001.
- (3) Martin, R. *Electronic Structure: Basic Theory and Practical Methods*; Cambridge University Press: Cambridge, U. K., 2004.
- (4) Fuchs, M.; Gonze, X. *Phys. Rev. B* **2002**, *65*, 235109.
- (5) Furche, F. *J. Chem. Phys.* **2008**, *129*, 114105.
- (6) Furche, F.; Van Voorhis, T. *J. Chem. Phys.* **2005**, *122*, 164106.
- (7) Janesko, B. G.; Henderson, T. M.; Scuseria, G. E. *J. Chem. Phys.* **2009**, *130*, 081105.
- (8) Toulouse, J.; Gerber, I. C.; Jansen, G.; Savin, A.; Angyan, J. G. *Phys. Rev. Lett.* **2009**, *102*, 096404.
- (9) Dion, M.; Rydberg, H.; Schroder, E.; Langreth, D. C.; Lundqvist, B. I. *Phys. Rev. Lett.* **2004**, *92*, 246401.
- (10) Lee, K.; Murray, E. D.; Kong, L. Z.; Lundqvist, B. I.; Langreth, D. C. *Phys. Rev. B* **2010**, *82*, 081101.
- (11) Vydrov, O. A.; Van Voorhis, T. *J. Chem. Phys.* **2010**, *133*, 244103.
- (12) Langreth, D. C.; Lundqvist, B. I.; Chakarova-Kack, S. D.; Cooper, V. R.; Dion, M.; Hyldgaard, P.; Kelkkanen, A.; Kleis, J.; Kong, L. Z.; Li, S.; Moses, P. G.; Murray, E.; Puzder, A.; Rydberg, H.; Schroder, E.; Thonhauser, T. *J. Phys., Condens. Matter* **2009**, *21*, 084203.
- (13) Tarnopolsky, A.; Karton, A.; Sertchook, R.; Vuzman, D.; Martin, J. M. L. *J. Phys. Chem. A* **2008**, *112*, 3.
- (14) Goerigk, L.; Grimme, S. *J. Chem. Theory Comput.* **2011**, *7*, 291.
- (15) Sharkas, K.; Toulouse, J.; Savin, A. *J. Chem. Phys.* **2011**, *134*, 064113.
- (16) Wu, X.; Vargas, M. C.; Nayak, S.; Lotrich, V.; Scoles, G. *J. Chem. Phys.* **2001**, *115*, 8748.
- (17) Wu, Q.; Yang, W. T. *J. Chem. Phys.* **2002**, *116*, 515.
- (18) Johnson, E. R.; Becke, A. D. *J. Chem. Phys.* **2005**, *123*, 024101.
- (19) Tao, J.; Perdew, J. P.; Ruzsinszky, A. *Phys. Rev. B* **2010**, *81*, 233102.
- (20) Grimme, S. *J. Comput. Chem.* **2006**, *27*, 1787.
- (21) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. *J. Chem. Phys.* **2010**, *132*, 154104.
- (22) Jurecka, P.; Cerny, J.; Hobza, P.; Salahub, D. R. *J. Comput. Chem.* **2007**, *28*, 555.
- (23) Riley, K. E.; Pitonak, M.; Jurecka, P.; Hobza, P. *Chem. Rev.* **2010**, *110*, 5023.
- (24) Sato, T.; Nakai, H. *J. Chem. Phys.* **2009**, *131*, 224104.
- (25) Tkatchenko, A.; Scheffler, M. *Phys. Rev. Lett.* **2009**, *102*, 073005.
- (26) Zhao, Y.; Truhlar, D. G. *Acc. Chem. Res.* **2008**, *41*, 157.
- (27) Zhao, Y.; Truhlar, D. G. *Theor. Chem. Acc.* **2008**, *120*, 215.
- (28) Kümmel, S.; Kronik, L. *Rev. Mod. Phys.* **2008**, *80*, 3.
- (29) Marom, N.; Bernstein, J.; Garel, J.; Tkatchenko, A.; Joselevich, E.; Kronik, L.; Hod, O. *Phys. Rev. Lett.* **2010**, *105*, 046801.
- (30) Marom, N.; Tkatchenko, A.; Scheffler, M.; Kronik, L. *J. Chem. Theory Comput.* **2010**, *6*, 81.
- (31) Tkatchenko, A.; Romaner, L.; Hofmann, O. T.; Zojer, E.; Ambrosch-Draxl, C.; Scheffler, M. *MRS Bull.* **2010**, *35*, 435.
- (32) Rossi, M.; Blum, V.; Kupser, P.; von Helden, G.; Bierau, F.; Pagel, K.; Meijer, G.; Scheffler, M. *J. Phys. Chem. Lett.* **2010**, *1*, 3465.
- (33) McNellis, E. R.; Meyer, J.; Reuter, K. *Phys. Rev. B* **2009**, *80*, 205414.
- (34) Tonigold, K.; Gross, A. J. *J. Chem. Phys.* **2010**, *132*, 224701.
- (35) Brede, J.; Atodiresei, N.; Kuck, S.; Lazic, P.; Caciuc, V.; Morikawa, Y.; Hoffmann, G.; Blugel, S.; Wiesendanger, R. *Phys. Rev. Lett.* **2010**, *105*, 047204.
- (36) Civalleri, B.; Zicovich-Wilson, C. M.; Valenzano, L.; Ugliengo, P. *CrystEngComm* **2008**, *10*, 405.
- (37) Civalleri, B.; Zicovich-Wilson, C. M.; Valenzano, L.; Ugliengo, P. *CrystEngComm* **2008**, *10*, 1693.
- (38) Rimola, A.; Civalleri, B.; Ugliengo, P. *Phys. Chem. Chem. Phys.* **2010**, *12*, 6357.
- (39) Fedorov, I. A.; Zhuravlev, Y. N.; Berveno, V. P. *Phys. Chem. Chem. Phys.* **2011**, *13*, 5679.
- (40) Bucko, T.; Hafner, J.; Lebegue, S.; Angyan, J. G. *J. Phys. Chem. A* **2010**, *114*, 11814.
- (41) Marom, N.; Tkatchenko, A.; Kapishnikov, S.; Kronik, L.; Leiserowitz, L. *Cryst. Growth Des.* **2011**, *11*, 3332.
- (42) Jenness, G. R.; Karalti, O.; Al-Saidi, W. A.; Jordan, K. D. *J. Phys. Chem. A* **2011**, *115*, 5955.
- (43) Vincent, M. A.; Hillier, I. H. *Phys. Chem. Chem. Phys.* **2011**, *13*, 4388.
- (44) Zhao, Y.; Truhlar, D. G. *Chem. Phys. Lett.* **2011**, *502*, 1.
- (45) Ferrighi, L.; Madsen, G. K. H.; Hammer, B. *Chem. Phys. Lett.* **2010**, *492*, 183.
- (46) Madsen, G. K. H.; Ferrighi, L.; Hammer, B. *J. Phys. Chem. Lett.* **2010**, *1*, 515.
- (47) Gu, J. D.; Wang, J.; Leszczynski, J.; Xie, Y. M.; Schaefer, H. F. *Chem. Phys. Lett.* **2008**, *459*, 164.
- (48) Raju, R. K.; Ramraj, A.; Hillier, I. H.; Vincent, M. A.; Burton, N. A. *Phys. Chem. Chem. Phys.* **2009**, *11*, 3411.
- (49) Bretherick, N. H.; van Mourik, T. *J. Chem. Theory Comput.* **2010**, *6*, 2687.
- (50) Hohenstein, E. G.; Chill, S. T.; Sherrill, C. D. *J. Chem. Theory Comput.* **2008**, *4*, 1996.
- (51) Ceperley, D. M.; Alder, B. J. *Phys. Rev. Lett.* **1980**, *45*, 566.
- (52) Perdew, J. P.; Wang, Y. *Phys. Rev. B* **1992**, *45*, 13244.
- (53) Vosko, S. H.; Wilk, L.; Nusair, M. *Can. J. Phys.* **1980**, *58*, 1200.
- (54) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865.
- (55) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1997**, *78*, 1396.
- (56) Perdew, J. P.; Ernzerhof, M.; Burke, K. *J. Chem. Phys.* **1996**, *105*, 9982.
- (57) Adamo, C.; Barone, V. *J. Chem. Phys.* **1999**, *110*, 6158.
- (58) Ernzerhof, M.; Scuseria, G. E. *J. Chem. Phys.* **1999**, *110*, 5029.
- (59) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648.
- (60) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. *J. Phys. Chem.* **1994**, *98*, 11623.
- (61) Heyd, J.; Scuseria, G. E.; Ernzerhof, M. *J. Chem. Phys.* **2006**, *124*, 219906.
- (62) Heyd, J.; Scuseria, G. E.; Ernzerhof, M. *J. Chem. Phys.* **2003**, *118*, 8207.
- (63) Tao, J.; Perdew, J. P.; Staroverov, V. N.; Scuseria, G. E. *Phys. Rev. Lett.* **2003**, *91*, 146401.
- (64) Perdew, J. P.; Kurth, S. In *A Primer in Density Functional Theory*; Fiolhais, C.; Nogueira, F.; Marques, M., Eds.; Springer: Berlin, 2003.
- (65) Karton, A.; Gruzman, D.; Martin, J. M. L. *J. Phys. Chem. A* **2009**, *113*, 8434.
- (66) Hirshfeld, F. L. *Theor. Chim. Acta* **1977**, *44*, 129.
- (67) Jurecka, P.; Sponer, J.; Cerny, J.; Hobza, P. *Phys. Chem. Chem. Phys.* **2006**, *8*, 1985.
- (68) Takatani, T.; Hohenstein, E. G.; Malagoli, M.; Marshall, M. S.; Sherrill, C. D. *J. Chem. Phys.* **2010**, *132*, 144104.
- (69) Blum, V.; Gehrke, R.; Hanke, F.; Havu, P.; Havu, V.; Ren, X.; Reuter, K.; Scheffler, M. *Comput. Phys. Commun.* **2009**, *180*, 2175.
- (70) Havu, V.; Blum, V.; Havu, P.; Scheffler, M. *J. Comput. Phys.* **2009**, *228*, 8367.
- (71) 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.; Montgomer, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, N. J.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, Ö.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. *Gaussian 09*, Revision A.1; Gaussian, Inc.: Wallingford, CT, 2009.

- (72) Řezáč, J.; Riley, K. E.; Hobza, P. *J. Chem. Theory Comput.* **2011**, *7*, 2427.
- (73) Vydrov, O. A.; Van Voorhis, T. *Phys. Rev. Lett.* **2009**, *103*, 063004.
- (74) Langreth, D. C.; Lundqvist, B. I. *Phys. Rev. Lett.* **2010**, *104*, 099303.
- (75) Vydrov, O. A.; Van Voorhis, T. *Phys. Rev. Lett.* **2010**, *104*, 099304.
- (76) Vydrov, O. A.; Van Voorhis, T. *Phys. Rev. A* **2010**, *81*, 062708.
- (77) Dobson, J. F.; McLennan, K.; Rubio, A.; Wang, J.; Gould, T.; Le, H. M.; Dinte, B. P. *Aust. J. Chem.* **2001**, *54*, 513.
- (78) Dobson, J. F.; White, A.; Rubio, A. *Phys. Rev. Lett.* **2006**, *96*, 073201.
- (79) Lebegue, S.; Harl, J.; Gould, T.; Angyan, J. G.; Kresse, G.; Dobson, J. F. *Phys. Rev. Lett.* **2010**, *105*, 196401.
- (80) Bjork, J.; Hanke, F.; Palma, C. A.; Samori, P.; Cecchini, M.; Persson, M. J. *Phys. Chem. Lett.* **2010**, *1*, 3407.
- (81) Arbuznikov, A. V.; Kaupp, M. *Chem. Phys. Lett.* **2003**, *381*, 495.
- (82) Marom, N.; Hod, O.; Scuseria, G. E.; Kronik, L. *J. Chem. Phys.* **2008**, *128*, 164107.
- (83) (a) Körzdörfer, T.; Kümmel, S.; Marom, N.; Kronik, L. *Phys. Rev. B* **2009**, *79*, 201205. (b) Körzdörfer, T.; Kümmel, S.; Marom, N.; Kronik, L. *Phys. Rev. B* **2010**, *82*, 129903.
- (84) Rissner, F.; Egger, D. A.; Natan, A.; Körzdörfer, T.; Kümmel, S.; Kronik, L.; Zojer, E. *J. Am. Chem. Soc.* **2011**, published online. DOI: 10.1021/ja203579c.
- (85) Marom, N.; Kronik, L. *Appl. Phys. A: Mater. Sci. Process.* **2009**, *95*, 159.
- (86) Dori, N.; Menon, M.; Kilian, L.; Sokolowski, M.; Kronik, L.; Umbach, E. *Phys. Rev. B* **2006**, *73*, 195208.
- (87) Bilc, D. I.; Orlando, R.; Shaltaf, R.; Rignanese, G. M.; Iniguez, J.; Ghosez, P. *Phys. Rev. B* **2008**, *77*, 165107.
- (88) Palumbo, M.; Hogan, C.; Sottile, F.; Bagala, P.; Rubio, A. *J. Chem. Phys.* **2009**, *131*, 084102.
- (89) Arantes, J. T.; Lima, M. P.; Fazzio, A.; Xiang, H.; Wei, S. H.; Dalpian, G. M. *J. Phys. Chem. B* **2009**, *113*, 5376.
- (90) Prodan, I. D.; Scuseria, G. E.; Martin, R. L. *Phys. Rev. B* **2006**, *73*, 045104.
- (91) Stroppa, A.; Kresse, G. *Phys. Rev. B* **2009**, *79*, 201201.
- (92) Varley, J. B.; Janotti, A.; Van de Walle, C. G. *Phys. Rev. B* **2010**, *81*, 245216.
- (93) Tkatchenko, A.; DiStasio, R. A., Jr; Car, R.; Scheffler, M. To be published.
- (94) Schramm, C. J.; Scaringe, R. P.; Stojakovic, D. R.; Hoffman, B. M.; Ibers, J. A.; Marks, T. J. *J. Am. Chem. Soc.* **1980**, *102*, 6702.
- (95) Solozhenko, V. L.; Will, G.; Elf, F. *Solid State Commun.* **1995**, *96*, 1.

Field-Extremum Model for Short-Range Contributions to Hydration Free Energy

Anna Pomogaeva and Daniel M. Chipman*

Radiation Laboratory, University of Notre Dame, Notre Dame, Indiana 46556-5674, United States

S Supporting Information

ABSTRACT: The performance in describing hydration free energies of a broad class of neutral, cationic, and anionic solutes is tested for the recently proposed FESR (Field-Extremum Short-Range) implicit solvation model for interactions between the solute and nearby water molecules, as taken in conjunction with the previously developed SS(V)PE (Surface and Simulation of Volume Polarization for Electrostatics) dielectric continuum model for long-range interactions with bulk water. The empirical FESR model mainly describes solute–water hydrogen bonding interactions by correlating them with the maximum and minimum values of the electric field produced by the solute at the surface of the cavity that excludes solvent. A preliminary report showed that, with only four adjustable parameters, the FESR model, in conjunction with SS(V)PE, can produce hydration energies comparable to the best analogous efforts in the literature that utilized many more parameters. Here, the performance of the FESR model is more fully documented in several respects. The dependence on the underlying quantum mechanical method used to treat the internal electronic structure of the solute is tested by comparing uncorrelated Hartree–Fock to correlated density functional calculations and by comparing a modest sized to a large basis set. The influence of cavity size is studied in connection with an isodensity contour construction of the cavity. The sensitivity of the results to the parameters in the FESR model is considered, and it is found that the dependence on the electric field strength is quite nonlinear, with an optimum exponent consistently in the range of 3 to 4. Overall, it is concluded that the FESR model shows considerable utility for improving the accuracy of implicit models of aqueous solvation.

INTRODUCTION

In the gas phase, it is now possible to accurately calculate properties and chemical reactions in many molecular systems using high level quantum mechanical (QM) methods. However, in condensed phases the environment impacts most chemical features so that such calculations in solution are still challenging and generally require a compromise between the theoretical reliability of an approach and its practical usefulness. Consequently, a wide variety of approaches has been developed to treat the effects of solvation. Direct attacks that explicitly treat many solvent molecules are capable of high accuracy but are computationally expensive. They remain expensive even if the solvent molecules are treated very approximately, as with molecular mechanical methods, due both to the necessity of including many solvent molecules to properly incorporate the long-range electrostatic solute–solvent interactions and to the need to carry out extensive statistical averaging over thermal motions in the solvent. Therefore, considerable interest remains in the development and improvement of highly simplified models that implicitly incorporate the most important influences of solvent on a solute. Several excellent reviews^{1–3} are available that comprehensively discuss implicit solvation models.

One particularly important property is the free energy of solvation, which governs solubility, thermodynamic stability, pK_a , redox potential, and many other chemical attributes of a solute. For polar solutes in polar solvents, the long-range electrostatic interactions between solvent and solute dominate the solvation energy, so it is first necessary to have as good a description of these interactions as possible. For this purpose, most implicit solvation approaches use a classical dielectric continuum

treatment based either on Poisson's equation⁴ or on a generalized Born approach⁵ to describe the statistically averaged response of bulk solvent.

Even with an accurate accounting of the long-range electrostatic interactions between solute and solvent, there generally remain significant short-range interactions that should be treated separately.^{6–21} These short-range interactions are usually phenomenologically divided into nonelectrostatic dispersion and exchange-repulsion energies and also the differences between the short-range electrostatic interactions that actually occur with nearby solvent molecules from those that are described by the dielectric continuum model. Additional short-range interactions such as cavitation work, charge transfer effects, three-body interactions, and nonlinear solvent response are also sometimes explicitly distinguished.^{1–3} Hydrogen bonding, while technically mostly arising from a combination of the aforementioned effects with perhaps also contributions from partial covalent character,^{22–30} is sufficiently unique that it is often considered a separate interaction.

In a preliminary report,³¹ we have described a very promising simple model having only a handful of parameters that implicitly describe some of the most important short-range effects of aqueous solvation. Used in conjunction with a modern dielectric continuum model of the long-range electrostatic effects, it provides quite reasonable hydration energies for a large collection of neutral and ionic solutes. In the present work, we further document this new model to show how it performs under different conditions of the calculation.

Received: August 16, 2011

Published: October 14, 2011

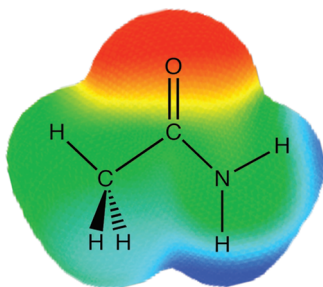


Figure 1. Acetamide molecule, colored according to the outgoing normal electric field F it generates on its $\rho_0 = 0.001$ au isodensity surface. The hottest red spot above the carbonyl oxygen is the position of F_{\min} , where acetamide is likely to accept a hydrogen bond from a solvent water molecule. The hottest blue spot below one of the amide hydrogens is the position of F_{\max} , where acetamide is likely to donate a hydrogen bond to a solvent water molecule.

Our preliminary report introduced the empirical FESR (Field Extremum Short-Range) model³¹ that was argued to mainly describe hydrogen-bonding effects on hydration energies. Earlier studies^{32–34} had indicated that the strength of a hydrogen bond depends on the electric field in its vicinity. In line with this, the FESR method basically searches the electric field produced by the solute at the cavity surface to find the hottest spots where hydrogen bonds can be expected to occur. Thus, an anion should accept a hydrogen bond from water at the position of minimum (i.e., most negative) outgoing normal electric field on the surface, while a cation formed by protonation of a neutral should donate a hydrogen bond to water at the position of maximum outgoing normal electric field on the surface. Neutral solutes should show similar although probably diminished effects. Furthermore, the strength of these hydrogen bonds should be related to the magnitude of the field at those positions. These ideas are illustrated in Figure 1.

The FESR model quantifies these ideas and provides a simple means with few parameters to describe short-range contributions to hydration energies. Together with a modest-level QM treatment of the solute and dielectric continuum treatment of long-range electrostatic contributions, the FESR model with only four adjustable parameters showed³¹ mean unsigned errors (MUE) from experiment of only 0.9 kcal/mol for 264 neutral solutes and 2.4 kcal/mol for 111 ionic solutes, which are comparable to the best analogous previous efforts in the literature with SM6 and SMVLE models that utilize many more parameters.²¹

The simplicity and accuracy of the FESR model makes it attractive for practical calculations, but the empirical nature of the approach demands proofs of its general reliability. In the present contribution, we explore the performance of the FESR model in several respects, including the influence of electron correlation and basis set size in the QM method used to obtain the internal electronic structure of the solute, the dependence on the size of the solute cavity, and the sensitivity to the adjustable parameters. Our development of this empirical model should also provide guidance into how a more fundamental implicit model of short-range solvation effects may be developed in future work.

METHOD AND COMPUTATIONAL DETAILS

Experimental Comparison. Testing the accuracy of the parametrization of any semiempirical model requires a pool of available experimental data. In the present research, experimental

hydration free energies were taken from the Minnesota Solvation Database, version 2009,³⁵ that has been collected and kindly provided as an open source database by the Minnesota group. The standard state assumed there and here corresponds to the Ben-Naim convention³⁶ of identical concentrations in gas and in solution. Ionic hydration free energies are based on the Tissandier et al.³⁷ value of -265.9 kcal/mol for the proton hydration free energy. The database developers estimate uncertainties in the experimental data to be about 0.2 kcal/mol for neutrals and about 3 kcal/mol for ions.³⁵

This database also includes optimized gas-phase geometries for all solutes. We adopted these geometries in our work as unchanged upon solvation. In principle, solvation energies will be affected by changes in geometry and zero-point vibration energy upon passing to the condensed phase,³⁸ but the solutes in this collection are not expected to undergo significant structural changes in solution. Therefore, it is believed that inclusion of such changes would not significantly affect our results at the level of accuracy being considered. Allowing the solutes to relax their geometries and change their vibrational frequencies in solution would presumably tend to further increase the accuracy of the results, although at the expense of sacrificing much of the simplicity of the approach.

This Minnesota Solvation Database contains 274 neutrals, 52 cations, and 60 anions, overall making 386 aqueous solutes. However, only 372 of them, including 261 neutrals, 51 cations, and 60 anions, were successfully treated by the single-center integration scheme used in our dielectric continuum program, which fails for severely nonspherical solutes. It should be noted that this is not a shortcoming of the model being presented but rather is only a limitation of the particular surface integration scheme implemented in the current version of the computer code that was utilized. The calculations failed either at the smallest cavity size or at one of the QM levels considered for 13 neutrals and one cation. Several of the solutes were successfully treated only after shifting the cavity surface integration center from the default origin at the center of nuclear charge to some other point. The full list of failed solutes and of those with a reset center of surface integration can be found for each computational case in Tables S1 and S2 of the Supporting Information. Note that our earlier report³¹ that used only one QM method and cavity size successfully treated three more neutrals for a total of 375 solutes. While the extent of the database used influences the empirical fitting of parameters, no significant differences exist between the results from there and here.

Quantum Mechanical Methods. Several combinations of QM methods were used to treat the internal electronic structure of the solute.

The effects of electron correlation in the solute were investigated by comparing uncorrelated Hartree–Fock (HF) *ab initio* methods with correlated B3LYP^{39–41} density functional methods.

The influence of the basis set was considered by comparing the modest-sized 6-31+G^{*42–44} basis with the rather bigger G3large⁴⁵ basis set. The former is double- ζ in the valence space and also has polarization and diffuse functions on all atoms but hydrogen. The latter is of triple- ζ quality in the valence space, has multiple polarization functions on all atoms, has diffuse functions on all atoms, and is fortunately available for all of the atoms included in the database. We remark that additional diffuse functions are not deemed necessary, in part because it has been found that application of a dielectric continuum tends to make anions smaller.^{46,47} In this connection, we note that application of a dielectric

continuum also tends to make cations larger,^{47,48} both of these trends being counterintuitive.

Dielectric Continuum Calculations. Long-range electrostatic contributions to the hydration free energy were calculated by means of the SS(V)PE (Surface and Simulation of Volume Polarization for Electrostatics) dielectric continuum model.^{49,50} This model pays particular attention to approximating the volume polarization arising from the long-range electrostatic effects of solute charge density that penetrates outside the cavity nominally enclosing it, the effects of which are treated more exactly in the SVPE (Surface and Volume Polarization for Electrostatics) model.^{51–55} If implemented with the same cavity, the SS(V)PE method becomes equivalent^{56,57} to the modified version of the IEF-PCM method that is often currently used.^{58,59}

To link a QM treatment of the solute with a classical dielectric continuum model of solvent, it is important to choose a proper physical boundary between the solute and the solvent, i.e., the cavity which excludes solvent. Unlike many other implementations that use atom-centered spheres to define the solute cavity, we prefer to use a cavity based on an electronic isodensity contour of the solute. This has the advantage of being naturally adapted to the solute shape and of requiring only a single parameter needed to determine the overall cavity size. For this work, we examined three different cavity sizes, corresponding to solute electronic isodensity contours of $\rho_0 = 0.0005, 0.001,$ and 0.002 au, which were found to be reasonable in previous studies.^{60,61} Surface integrations over the cavity were carried out by single-center Lebedev quadrature with 1202 grid points. The water dielectric constant was set to 78.36.

Short-Range Contributions. After determination of $\Delta G_{\text{SS(V)PE}}$ for the dielectric continuum contributions to hydration free energies, an additional correction term for short-range contributions is calculated by means of the FESR method.³¹

$$\Delta G_{\text{exptl}} \approx \Delta G_{\text{SS(V)PE}} + \Delta G_{\text{FESR}} \quad (1)$$

Here, ΔG_{FESR} is an empirically parametrized term depending on the minimum and maximum values of the outgoing normal electric field produced by the solute anywhere on the cavity surface.

$$\Delta G_{\text{FESR}} = a + b|F_{\text{min}}|^d + cF_{\text{max}}^d \quad (2)$$

The value of F_{min} is taken as the lower of zero or the most negative value of the outgoing normal electric field, and F_{max} is the higher of zero or the most positive value of the outgoing normal electric field. The empirical parameters $a, b, c,$ and d are adjusted to minimize the least-squares deviation from experimental values. Their values are reported here to yield free energy in kilocalories per mole when the electric fields are expressed in atomic units. Unless specified otherwise, equal weight is given to all solutes in the training set. In some particular cases, we also tested a fitting strategy that utilized different weights according to the estimated experimental error for each solute.

As with many other treatments of short-range effects,^{62–64} the approach of eq 2 only includes the short-range interactions as a sum of independently calculated additive contributions to the solvation energy. Ideally, the short-range interactions should be allowed to polarize the electronic structure of the solute by including them as potential energy terms in the solute Hamiltonian,^{65–67} in a manner similar to the way the long-range dielectric effects are usually included. One of our eventual goals is to adapt the FESR approach to produce such a potential energy term in the solute Hamiltonian in order to treat the influence of short-range

solute–solvent interactions in a self-consistent reaction field framework.

The fitting of parameters in the FESR model was done for all 12 combinations from each of the three cavity sizes with $\rho_0 = 0.0005, 0.001,$ and 0.002 au contours together with each of the four QM levels corresponding to HF/6-31+G*, B3LYP/6-31+G*, HF/G3large, and B3LYP/G3large.

All of the computations were made using a locally modified version of the GAMESS software.⁶⁸

RESULTS AND DISCUSSION

General Remarks. For the most part, results will be discussed only for broad groups, such as the class of all solutes taken together or for the individual subclasses of neutrals, cations, and anions. For reference purposes, Tables S3–S6 of the Supporting Information give full details of the electric field extrema on the cavity surface, $\Delta G_{\text{SS(V)PE}}, \Delta G_{\text{FESR}},$ and the related errors for each individual solute from the various combinations of computational methods considered.

The overall mean unsigned errors (MUE) for calculated hydration free energies compared to experiment are given in Table 1 for various combinations of QM methods and cavity sizes. Uncorrected SS(V)PE results are also given for reference purposes. The SS(V)PE results generally show the smallest errors for neutrals, somewhat larger errors for cations, and very large errors for anions. The SS(V)PE+FESR results are shown first for allowing the d parameter to optimize separately for each individual QM method and then for constraining the d parameter to have a common universal value for all of the QM methods. It is immediately evident that in all cases the FESR correction substantially reduces the errors over SS(V)PE alone, most particularly for anions.

The SS(V)PE+FESR MUE results for neutrals are generally somewhat larger than the estimated experimental errors of about 0.2 kcal/mol, the MUE results for cations are either comparable to or in some instances significantly lower than the estimated experimental errors of about 3 kcal/mol, and the MUE results for anions are either comparable to and in some instances slightly lower than the estimated experimental errors of about 3 kcal/mol. The most obvious explanation for how some MUE results can be lower than the estimated experimental errors is that the latter estimates are actually too high. Several other possible explanations of this matter have been previously elaborated.³¹

The results in Table 1 derive from giving equal weights to all solutes in the fitting process. Allowing different weights for different classes of solutes according to their estimated experimental uncertainties (i.e., relative weights of 1/0.2 for neutrals and 1/3 for ions) leads to quite similar results. For example, for B3LYP/6-31+G* with $\rho_0 = 0.001$ au, the SS(V)PE+FESR MUE from different weighting is 1.40 kcal/mol for all solutes, as compared to 1.37 kcal/mol obtained from equal weighting.

A convenient visualization of these results is given in Figure 2 for the particular case of B3LYP/6-31+G* calculations with $\rho_0 = 0.001$ au. All of the other various computational combinations considered also show quite similar behavior. The results in Figure 2 would all fall on the diagonal solid lines if perfect agreement with experimental results were achieved. The top left panel shows hydration free energies calculated from the SS(V)PE model alone compared to experimental results for all solutes, while the top right panel shows an expanded version for the neutral solutes. The agreement with experimental results is seen

Table 1. Mean Unsigned Errors in kcal/mol Obtained from Several QM Methods and Cavity Sizes^a

solvation method	parameter <i>d</i> status ^b	all solutes (372)	neutrals (261)	cations (51)	anions (60)
HF/6-31+G*					
SS(V)PE		4.65, 3.92, 5.28	1.65, 2.56, 4.63	7.13, 2.28, 6.82	15.61, 11.21, 6.77
SS(V)PE+FESR	individual <i>d</i>	1.73, 1.54, 2.03	1.19, 1.16, 1.67	3.24, 2.14, 3.11	2.80, 2.69, 2.66
SS(V)PE+FESR	universal <i>d</i>	1.76, 1.58, 2.03	1.26, 1.24, 1.69	3.05, 2.14, 3.07	2.83, 2.58, 2.64
B3LYP/6-31+G*					
SS(V)PE		5.19, 3.98, 4.85	1.51, 1.81, 3.45	8.37, 2.95, 5.91	18.50, 14.32, 10.07
SS(V)PE+FESR	individual <i>d</i>	1.65, 1.37, 1.71	1.09, 0.94, 1.15	2.99, 1.80, 2.95	2.97, 2.88, 3.12
SS(V)PE+FESR	universal <i>d</i>	1.65, 1.37, 1.71	1.09, 0.94, 1.15	2.99, 1.80, 2.95	2.97, 2.88, 3.12
HF/G3large					
SS(V)PE		4.86, 3.65, 4.65	1.63, 1.94, 3.68	8.00, 2.80, 6.43	16.22, 11.83, 7.33
SS(V)PE+FESR	individual <i>d</i>	1.93, 1.71, 2.15	1.52, 1.45, 1.83	2.99, 2.01, 3.29	2.81, 2.57, 2.61
SS(V)PE+FESR	universal <i>d</i>	1.93, 1.71, 2.15	1.52, 1.45, 1.84	2.99, 2.01, 3.26	2.81, 2.57, 2.60
B3LYP/G3large					
SS(V)PE		5.43, 4.07, 4.52	1.61, 1.66, 3.01	9.26, 3.87, 5.20	18.80, 14.69, 10.50
SS(V)PE+FESR	individual <i>d</i>	1.71, 1.45, 1.66	1.15, 1.00, 1.11	3.01, 1.97, 2.69	3.04, 2.95, 3.17
SS(V)PE+FESR	universal <i>d</i>	1.73, 1.48, 1.67	1.14, 1.00, 1.09	3.18, 2.04, 2.74	3.05, 3.06, 3.27

^aThe three numbers in each slot correspond to results for cavity sizes from $\rho_0 = 0.0005, 0.001, \text{ and } 0.002 \text{ au}$, respectively. ^b“Individual *d*” means the parameter *d* is optimized individually for each QM method. “Universal *d*” means the parameter *d* is constrained to have a common universal value for all of the QM methods.

to be reasonable for neutrals and cations, while large and systematic errors are found for anions. The bottom panels show analogous results for the SS(V)PE+FESR model. Some notable improvement is found for neutrals and cations, and dramatic improvement is obtained for anions.

It is generally believed that solute surface area and/or volume are significant descriptors of generic short-range interactions between the solute and solvent.^{69–77} However, our computational experiments (data not shown) found that further inclusion of terms proportional to solute surface area or volume led to only tiny improvements in the MUE. We surmise that the constant term *a* in the FESR model likely describes in some rough overall average sense such size-dependent contributions. It may then be that the solutes in the database used for training do not cover a sufficient range of sizes to specifically observe beyond this the influence of solute size over whatever other errors remain. In this connection, it may be noted that studies of cavitation have shown that in addition to size, solute shape is also a significant complicating factor.^{78–80}

While it is believed that the field-dependent terms in the FESR model mainly describe hydrogen bonding effects, it should be pointed out that the empirical optimization of parameters necessarily also implicitly incorporates some contributions from dispersion and exchange-repulsion interactions in the model.

For some solutes, there are several hot spots where hydrogen bonds might be expected to occur. The FESR method takes care of such cases only insofar as the accumulated effects of all hydrogen bonds that are incorporated in the experimental values affect the values of the fitting parameters. A preliminary attempt to include more than one contribution to ΔG_{FESR} for some representative solutes having more than one local region of minimum and/or maximum electric field actually led to poorer agreement with experimental results. While this was done without a reoptimization of parameters, it is anyway clear that this matter is more subtle than it seems and will require further research to understand properly.

As a control experiment, we also tested the possibility that the FESR model might perform well alone, without any contributions from a dielectric continuum model. Even in the best case, this led to a very large MUE for neutrals of 9 kcal/mol and for ions of 20 kcal/mol, thus soundly refuting the possibility.

Dependence on QM Method. The QM method used to determine the electronic structure of the solute affects the hydration energy calculated in this work in two distinct ways. First, the QM method determines the solute size and shape through specifying the isodensity surface used to define the cavity. In this connection, we note that inclusion of electron correlation typically changes the cavity size by only about 0.01 Å (usually, but not always, increasing it), which is very small relative to the changes in cavity size of typically about 0.2 Å in passing from a ρ_0 of 0.0005 to one of 0.001 or from 0.001 to 0.002 au. Extension of the basis set has even less of an effect on the cavity size, usually slightly reducing it by less than 0.001 Å with HF and less than 0.004 Å with B3LYP. Therefore, any trends in results from different QM methods that are consistent at all cavity sizes examined are not due to the differences in how the various QM methods determine the cavity. Second, the QM method determines the electric properties presented by the solute at the cavity surface, specifically the electrostatic potential that is utilized in the SS(V)PE calculation and the electric field that is utilized in the FESR correction, and this is then the main route through which the QM method influences the results.

It is seen in Table 1 that each of the QM methods considered leads to good overall results. Even so, there are some notable small differences in performance that are attributable to electron correlation and basis set effects.

Inclusion of electron correlation, as done here by comparison of B3LYP to HF, should be expected to improve the results. With both basis sets, it is found that inclusion of electron correlation does indeed improve the SS(V)PE+FESR agreement with experimental results.

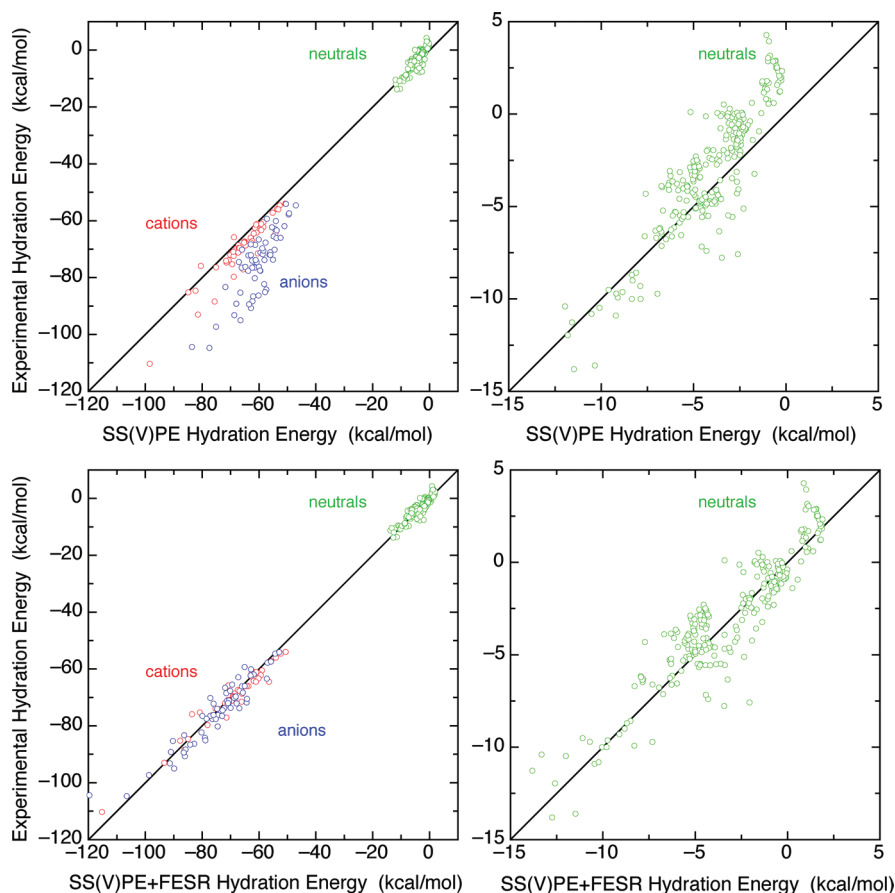


Figure 2. Comparison of experimental hydration energies with those calculated from the B3LYP/6-31+G* method with the $\rho_0 = 0.001$ au cavity size. The left panels show SS(V)PE and SS(V)PE+FESR results for all solutes, respectively, while the right panels show expanded versions covering only neutral solutes.

Extension of the basis set, as done here by comparison of G3large to 6-31+G*, should also be expected to improve the results. But with both HF and B3LYP, it is found that basis set extension instead slightly degrades the agreement with experimental results. A possible explanation for this apparently anomalous behavior of basis set extension is that the modest 6-31+G* basis set may accidentally tend to give better solute electric properties in the cavity surface region than does the larger G3large basis set. There is clearly a need for further research on this point to determine if other large basis sets give similar behavior.

Dependence on FESR Fitting Parameters. Examination of the optimum values of the parameters a , b , c , and d reported for all computational combinations in Table 2 shows that the value of b is always more than an order of magnitude larger than that of c . This is consistent with the substantially larger errors from SS(V)PE itself for anions as compared to cations or neutrals, which therefore requires larger FESR corrections for anions.

As expected on physical grounds discussed in the Introduction and seen from the detailed results in Tables S3–S6 of the Supporting Information, the FESR term involving F_{\min} makes its largest contribution to anions; the term involving F_{\max} makes its largest contribution to cations. Both contribute significantly to polar neutrals. In addition to the field-dependent terms, the FESR model also includes a constant term a which likely describes contributions from cavitation, exchange repulsion, and dispersion in some rough overall average sense.

Examples of the sensitivity of the results to d are shown in Figure 3, which correspond to the case of d being optimized separately for each individual QM method, while a , b , and c are reoptimized for each value of d . The top panel gives B3LYP/6-31+G* results for each cavity size as a function of d , showing that changes on the order of one unit in d lead to a significant change in the MUE. The bottom panel of Figure 3 gives results from different QM methods for the $\rho_0 = 0.001$ au cavity size as a function of d and again shows a similar sensitivity to d .

Interestingly, experiments allowing F_{\min} and F_{\max} to have different values of d (data not shown) led to negligible improvement in the overall MUE. This implies that both cations and anions prefer the same value of d , suggesting it has some underlying physical significance.

It is seen in Table 2 that the optimum d varies only from 2.9 to 4.1 for all of the combinations of methods considered, and over an even smaller range for the various QM methods at each particular cavity size. This suggests that it might be provident to claim a universal value of d that is independent of the QM method, although still dependent on the cavity size and separate optimization for each QM method of the a , b , and c parameters. Table 1 includes the results from assuming that all of the QM methods share a common universal value of d , where it is seen that this constraint sacrifices less than (and usually much less than) 0.10 kcal/mol in MUE over allowing each QM method to have its own optimum d . The optimum parameter values under

Table 2. Optimum Parameter Values Obtained from Several QM Methods and Cavity Sizes, Allowing Each QM Method to Have Its Own Individual Optimum d at Each Cavity Size

parameter	$\rho_0 = 0.0005$ au	$\rho_0 = 0.001$ au	$\rho_0 = 0.002$ au
HF/6-31+G*			
a	1.850	2.828	5.238
b	-0.378×10^6	-1.319×10^6	-0.491×10^6
c	-2.029×10^4	-1.897×10^4	0.066×10^4
d	3.4	4.1	4.0
B3LYP/6-31+G*			
a	1.570	2.345	4.158
b	-0.259×10^6	-0.585×10^6	-0.523×10^6
c	-1.543×10^4	-1.107×10^4	0.027×10^4
d	3.1	3.6	3.8
HF/G3large			
a	1.157	2.094	4.251
b	-0.168×10^6	-0.362×10^6	-0.403×10^6
c	-1.193×10^4	-0.687×10^4	0.045×10^4
d	3.1	3.6	3.9
B3LYP/G3large			
a	1.488	2.122	3.734
b	-0.145×10^6	-0.258×10^6	-0.318×10^6
c	-1.174×10^4	-0.674×10^4	-0.025×10^4
d	2.9	3.3	3.6

this constraint are given in Table 3. Comparison to Table 2 shows that the a , b , and c parameters become somewhat different under this constraint, changing in some instances by more than a factor of 2. This explains the failure of an attempt to find universal values of all four parameters for all of the QM methods considered (data not shown), which led to very large MUEs at each cavity size.

The results reported in the tables correspond to optimizing the FESR parameters over the class of all solutes taken together. At the expense of losing generality, even better results could be obtained if the parameters were to be optimized separately for each of the subclasses of neutrals, cations, and anions. For example, with the B3LYP/6-31+G* method taken with the $\rho_0 = 0.001$ au cavity size, such separate optimizations lead to MUE of 0.85, 1.61, and 2.70 kcal/mol for neutrals, cations, and anions, respectively. This corresponds to improvements of about 0.1–0.2 kcal/mol for these subclasses over the results in Table 1, where the parameters take on more compromise values from optimization over all classes.

To get an idea of the extent to which the parameters and results might depend on the database, the 372 solutes were divided into two groups by alternately assigning each solute in the given list to either a training set or a test set. B3LYP/6-31+G* calculations with the $\rho_0 = 0.001$ au cavity size together with optimization of parameters only over solutes in the training set led to a change in the MUE of the test set by only 0.02 kcal/mol from that of training over the entire database. The fitting parameters a , b , and c changed by less than 10%, and d changed not at all. A second analogous test made by interchanging the roles of the training and test sets produced essentially the same results. This exercise demonstrates the stability and robustness of the parametrization and further indicates that the model should

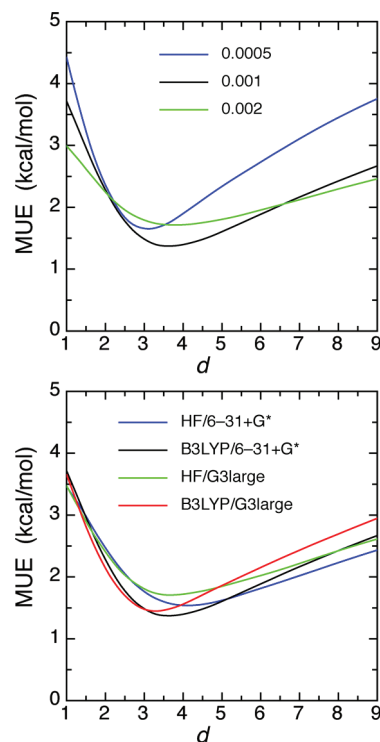


Figure 3. Mean unsigned error as a function of exponent d . The top panel shows results for the B3LYP/6-31+G* method with several cavity sizes. The bottom panel shows results for the $\rho_0 = 0.001$ au cavity size with several QM methods.

have a useful predictive power for solvation energies of molecules not contained in the data set.

Dependence on Cavity Size. Most of the trends with cavity size have already been discussed above in connection with other related matters. It remains here to explicitly note that the best SS(V)PE+FESR MUE results are obtained in all cases examined with the contour $\rho_0 = 0.001$ au, which is the most commonly used value and has also proved to be near optimal in previous studies of dielectric effects in neutrals.^{60,61} In contrast, with SS(V)PE alone, the lowest MUE for neutrals occurs with $\rho = 0.0005$ au, for cations with $\rho = 0.001$ au, and for anions with $\rho = 0.002$ au.

Among the various classes of solutes, cations are the most sensitive to the cavity size. The change in SS(V)PE+FESR MUE with each step in cavity size is about 1 kcal/mol for cations, whereas this change is about 0.3 kcal/mol for neutrals and about 0.2 kcal/mol for anions.

Behavior of Selected Solutes. Here, we describe some particularly notable behaviors of certain individual solutes, mainly to indicate areas where additional improvements can be sought in future work. To keep the discussion manageable, we discuss only the case of B3LYP/6-31+G* calculations with the $\rho_0 = 0.001$ au contour and optimization of FESR parameters over all classes of solutes. Other computational protocols show similar behavior.

The worst numerical result among the neutrals is for N,N'-dimethylpiperazine, where the experimental hydration energy is -7.58 kcal/mol while the SS(V)PE+FESR result is -2.17 kcal/mol, corresponding to an error of 5.41 kcal/mol. The FESR result in this case is actually slightly worse than the SS(V)PE result, which is -2.86 kcal/mol, corresponding to an error of 4.72 kcal/mol for this compound.

Table 3. Optimum Parameter Values Obtained from Several QM Methods and Cavity Sizes, under the Constraint of All QM Methods Sharing a Common Universal Value of d at Each Cavity Size

parameter	$\rho_0 = 0.0005$ au	$\rho_0 = 0.001$ au	$\rho_0 = 0.002$ au
d	3.1	3.6	3.8
HF/6-31+G*			
a	2.283	3.221	5.308
b	-0.162×10^6	-0.343×10^6	-0.293×10^6
c	-1.101×10^4	-0.756×10^4	0.055×10^4
B3LYP/6-31+G*			
a	1.570	2.345	4.158
b	-0.259×10^6	-0.585×10^6	-0.523×10^6
c	-1.543×10^4	-1.107×10^4	0.027×10^4
HF/G3large			
a	1.157	2.094	4.284
b	-0.168×10^6	-0.362×10^6	-0.311×10^6
c	-1.193×10^4	-0.687×10^4	0.041×10^4
B3LYP/G3large			
a	1.119	1.788	3.623
b	-0.260×10^6	-0.598×10^6	-0.544×10^6
c	-1.794×10^4	-1.171×10^4	-0.036×10^4

Being quite nonpolar, the 41 neutral unsubstituted hydrocarbons in the database generally have very small experimental hydration energies, in some cases even being positive, indicating hydrophobicity. By itself, SS(V)PE necessarily produces negative hydration energies, and it therefore gives a poor description of these solutes. The electric fields on the cavity surface of such solutes are generally quite small, consistent with essentially no hydrogen bonding to water, so the field-dependent terms in the FESR model are also ineffective in describing these solutes. In particular, the field-dependent terms contribute only about -0.01 kcal/mol for alkanes, about -0.07 kcal/mol for alkenes, about -0.25 kcal/mol for alkynes, and about -0.02 kcal/mol for aromatic hydrocarbons. The a term in the FESR model then becomes the most significant contribution, and its positive value is mainly responsible for the improvement in description of the unsubstituted hydrocarbons. The neutral unsubstituted hydrocarbons are all clustered in the top right corners of the panels in Figure 2, where it is seen that the FESR correction does not noticeably change the pattern for these solutes, except for providing a constant horizontal shift from the a term. Similar conclusions regarding an analogous positive constant term were drawn in a previous study of alkane hydration energies.⁸¹

For neutral halocarbons, the field dependent terms provide free energy contributions ranging from -0.03 to -0.37 kcal/mol, such that the total FESR correction is 1.97 to 2.31 kcal/mol. The large range in experimental free energies for this class of compounds (from -2.73 kcal/mol for bromotoluene up to 4.28 kcal/mol for octafluoropropane) is already mostly accounted for in the SS(V)PE results, while the constant shift from the FESR a term helps to reduce the MUE for this subclass from 2.38 to 0.68 kcal/mol.

Among cations, the errors in hydration energies of protonated 4-nitroaniline, 3-aminoaniline, ammonia, and dimethyl sulfide are increased by more than 1 kcal/mol upon FESR correction, and the errors are found to be quite sensitive to computation

level and cavity setting. All other cation results are improved by the FESR correction.

All anions improve their hydration energies upon applying the FESR correction. The largest anion error occurs for F^- , not surprisingly in light of its extremely high concentration of negative charge into a very small volume, where the error is reduced from the huge -21.3 kcal/mol at the SS(V)PE level to the still large 13.4 kcal/mol after FESR correction. However, in the case of OH^- , which is the other very small anion in the database, the similarly huge SS(V)PE error of -26.8 kcal/mol is quite effectively reduced by the FESR correction to only 3.7 kcal/mol.

CONCLUDING REMARKS

The SS(V)PE+FESR model to calculate hydration free energy was introduced in a previous short communication³¹ that demonstrated its potential utility and efficiency. The present contribution studies the performance of this model under a variety of computational conditions and also gives additional information on its behavior in various respects.

While all of the QM methods considered give good performance in this application, there are some differences worthy of note. Inclusion of electron correlation in the solute QM method, as studied by comparison of B3LYP to HF calculations, leads overall to better results. However, extension of the basis set in the solute QM method, as studied by comparison of the G3large to the modest 6-31+G* basis set, generally leads to slightly poorer results. The cavity obtained from the $\rho_0 = 0.001$ au contour gives significantly better results than with either larger or smaller cavities from 0.0005 or 0.002 au contours, respectively.

Overall, the best SS(V)PE+FESR results in this work are obtained with the B3LYP/6-31+G* QM method and the 0.001 au cavity contour. This gives a MUE for neutrals of 0.9 kcal/mol, for cations of 1.8 kcal/mol, for anions of 2.9 kcal/mol, and, combining the latter two, for all ions of 2.4 kcal/mol. This performance is comparable to the best analogous literature results using the SMVLE method²¹ that was trained over almost the same database as used here and that used the SVPE dielectric continuum model, which should give essentially the same results as SS(V)PE for the solutes concerned. After optimization of a large number of parameters, including 20 atom-specific surface tension coefficients and 12 generic parameters involving the solute electric field on the cavity surface, the SMVLE model achieved a MUE of 0.5 kcal/mol for the neutrals and 3.1 kcal/mol for the ions. Compared to this, the best SS(V)PE+FESR results are not quite as good for neutrals and are better for ions, despite the much smaller number of adjustable parameters. Our view is that minimizing the number of empirical parameters in a model increases the probability that the parameters will have physical significance that can be used to guide development of more fundamentally based theories.

The FESR correction utilizes a functional form where hydration free energies have a nonlinear dependence on the minimum and maximum values of the solute normal electric field on the cavity surface. It is argued that these field-dependent terms mainly describe specific hydrogen bonding interactions between solute and solvent. The nonlinear parameter d of the model is always found to be in the range of 2.9–4.1 with all of the combinations of computational protocols considered, its more exact value depending to some extent on QM method and cavity size. In fact, nearly universal values of d exist for each cavity size independent of the QM method used. Furthermore, negligible

improvement in the MUE is achieved by allowing cations and anions to have different values of d . These observations suggest some real underlying physical significance of the nonlinear d exponent of the field extrema for describing hydrogen bonding energies.

The SS(V)PE+FESR model also invokes a constant term a that significantly improves the hydration energies of nonpolar solutes and, by extension, of other solutes as well. This parameter likely describes in some rough overall average sense the short-range contributions from cavitation, exchange repulsion, and dispersion. While such interactions are actually size-dependent, no significant effect was found on including additional terms proportional to solute surface area or volume. This may be due to an insufficient range of sizes in the database used to train the parameters.

Further research will focus on more specifically modeling short-range contributions from cavitation, exchange repulsion, and dispersion, which can be expected to particularly improve the treatment of neutral solutes. Improvement in the treatment of ions should also follow, but that may be difficult to discern in practice because the best SS(V)PE+FESR results for ions are already in better agreement with experimental results than the estimated experimental error.

It is concluded that the FESR correction is quite effective in removing a large part of the error in hydration energies from the SS(V)PE model. There is every reason to believe that an analogous FESR correction would perform similarly well as an adjunct to other dielectric continuum models in common use. With the parameters reported in this work, the SS(V)PE+FESR model can be used to predict hydration energies of solutes not in the database used for its training. Additionally, it should also serve as a valuable building block for further development of models to bring results from implicit solvation calculations into even closer agreement with experimental results.

■ ASSOCIATED CONTENT

S Supporting Information. Tables of compounds that failed the surface integration for at least one QM protocol (S1); compounds that required a nondefault surface integration origin (S2); and listings for each solute under each QM protocol of the electric field extrema, hydration free energies, and related errors (S3, S4, S5, S6). This material is available free of charge via the Internet at <http://pubs.acs.org.proxy.library.nd.edu>.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: chipman.1@nd.edu.

■ ACKNOWLEDGMENT

The generosity of the Minnesota group in making available their extensive Solvation Database is gratefully acknowledged. This material is based upon work supported by the Department of Energy under Award Number DE-SC0002216. This is Contribution No. NDRL-4897 from the Notre Dame Radiation Laboratory.

■ REFERENCES

- (1) Tomasi, J.; Persico, M. *Chem. Rev.* **1994**, *94*, 2027.
- (2) Cramer, C. J.; Truhlar, D. G. *Chem. Rev.* **1999**, *99*, 2161.
- (3) Tomasi, J.; Mennucci, B.; Cammi, R. *Chem. Rev.* **2005**, *105*, 2999.

- (4) Miertuš, S.; Scrocco, E.; Tomasi, J. *Chem. Phys.* **1981**, *55*, 117.
- (5) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. *J. Am. Chem. Soc.* **1990**, *112*, 6127.
- (6) Sitkoff, D.; Sharp, K. A.; Honig, B. *J. Phys. Chem.* **1994**, *98*, 1978.
- (7) Marten, B.; Kim, K.; Cortis, C.; Friesner, R. A.; Murphy, R. B.; Ringnalda, M. N.; Sitkoff, D.; Honig, B. *J. Phys. Chem.* **1996**, *100*, 11775.
- (8) Barone, V.; Cossi, M.; Tomasi, J. *J. Chem. Phys.* **1997**, *107*, 3210.
- (9) Klamt, A.; Jonas, V.; Bürger, T.; Lohrenz, J. C. W. *J. Phys. Chem. A* **1998**, *102*, 5074.
- (10) Amovilli, C.; Barone, V.; Cammi, R.; Cancès, E.; Cossi, M.; Mennucci, B.; Pomelli, C.; Tomasi, J. *Adv. Quant. Chem.* **1999**, *32*, 227.
- (11) Pliego, J. R.; Riveros, J. M. *J. Phys. Chem. A* **2001**, *105*, 7241.
- (12) Kelly, C. P.; Cramer, C. J.; Truhlar, D. G. *J. Chem. Theory Comput.* **2005**, *1*, 1133.
- (13) Piquemal, J. P.; Marquez, A.; Parisel, O.; Giessner-Prettre, C. *J. Comput. Chem.* **2005**, *26*, 1052.
- (14) Kelly, C. P.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. B* **2006**, *110*, 16066.
- (15) Rizzo, R. C.; Aynechi, T.; Case, D. A.; Kuntz, I. D. *J. Chem. Theory Comput.* **2006**, *2*, 128.
- (16) Marenich, A. V.; Olson, R. M.; Kelly, C. P.; Cramer, C. J.; Truhlar, D. G. *J. Chem. Theory Comput.* **2007**, *3*, 2011.
- (17) Cramer, C. J.; Truhlar, D. G. *Acc. Chem. Res.* **2008**, *41*, 760.
- (18) Klamt, A.; Mennucci, B.; Tomasi, J.; Barone, V.; Curutchet, C.; Orozco, M.; Luque, F. J. *Acc. Chem. Res.* **2009**, *42*, 489.
- (19) Cramer, C. J.; Truhlar, D. G. *Acc. Chem. Res.* **2009**, *42*, 493.
- (20) Zuo, C.-S.; Wiest, O.; Wu, Y.-D. *J. Phys. Chem. A* **2009**, *113*, 12028.
- (21) Liu, J.; Kelly, C. P.; Goren, A. C.; Marenich, A. V.; Cramer, C. J.; Truhlar, D. G.; Zhan, C.-G. *J. Chem. Theory Comput.* **2010**, *6*, 1109.
- (22) Umeyama, H.; Morokuma, K. *J. Am. Chem. Soc.* **1977**, *99*, 1316.
- (23) Reed, A. E.; Weinhold, F. *J. Chem. Phys.* **1983**, *78*, 4066.
- (24) Stevens, W. J.; Fink, W. H. *Chem. Phys. Lett.* **1987**, *139*, 15.
- (25) Glendening, E. D.; Streitwieser, A. *J. Chem. Phys.* **1994**, *100*, 2900.
- (26) Chen, W.; Gordon, M. S. *J. Phys. Chem.* **1996**, *100*, 14316.
- (27) Isaacs, E. D.; Shukla, A.; Platzman, P. M.; Hamann, D. R.; Barbiellini, B.; Tulk, C. A. *Phys. Rev. Lett.* **1999**, *82*, 600.
- (28) Ghanty, T. K.; Staroverov, V. N.; Koren, P. R.; Davidson, E. R. *J. Am. Chem. Soc.* **2000**, *122*, 1210.
- (29) Barbiellini, B.; Shukla, A. *Phys. Rev. B* **2002**, *66*, 235101.
- (30) Khaliullin, R. Z.; Bell, A. T.; Head-Gordon, M. *Chem.—Eur. J.* **2009**, *15*, 851.
- (31) Pomogaeva, A.; Thompson, D. W.; Chipman, D. M. *Chem. Phys. Lett.* **2011**, *511*, 161.
- (32) Sadlej, J.; Buch, V.; Kazimirski, J. K.; Buck, U. *J. Phys. Chem. A* **1999**, *103*, 4933.
- (33) Chipman, D. M. *J. Chem. Phys.* **2003**, *118*, 9937.
- (34) Chipman, D. M.; Chen, F. W. *J. Chem. Phys.* **2006**, *124*, 144507.
- (35) Marenich, A. V.; Kelly, C. P.; Thompson, J. D.; Hawkins, G. D.; Chambers, C. C.; Giesen, D. J.; Winget, P.; Cramer, C. J.; Truhlar, D. G. *Minnesota Solvation Database*, version 2009; University of Minnesota: Minneapolis, MN, 2009.
- (36) Ben-Naim, A. *J. Phys. Chem.* **1978**, *82*, 792.
- (37) Tissandier, M. D.; Cowen, K. A.; Feng, W. Y.; Gundlach, E.; Cohen, M. H.; Earhart, A. D.; Coe, J. V.; Tuttle, T. R. *J. Phys. Chem. A* **1998**, *102*, 7787.
- (38) Ho, J.; Klamt, A.; Coote, M. L. *J. Phys. Chem. A* **2010**, *114*, 13442.
- (39) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648.
- (40) Lee, C. T.; Yang, W. T.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785.
- (41) We use the B3LYP version implementing the VWN5 electron gas correlation formula.
- (42) Hehre, W. J.; Ditchfield, R.; Pople, J. A. *J. Chem. Phys.* **1972**, *56*, 2257.
- (43) Hariharan, P. C.; Pople, J. A. *Theor. Chim. Acta* **1973**, *28*, 213.
- (44) Clark, T.; Chandrasekhar, J.; Spitznagel, G. W.; Schleyer, P. V. *J. Comput. Chem.* **1983**, *4*, 294.

- (45) Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Rassolov, V.; Pople, J. A. *J. Chem. Phys.* **1998**, *109*, 7764.
- (46) Luque, F. J.; Orozco, M.; Bhadane, P. K.; Gadre, S. R. *J. Chem. Phys.* **1994**, *100*, 6718.
- (47) Cabral do Couto, P. C.; Chipman, D. M. *J. Phys. Chem. A* **2010**, *114*, 12788.
- (48) Luque, F. J.; Gadre, S. R.; Bhadane, P. K.; Orozco, M. *Chem. Phys. Lett.* **1995**, *232*, 509.
- (49) Chipman, D. M. *J. Chem. Phys.* **1999**, *110*, 8012.
- (50) Chipman, D. M. *J. Chem. Phys.* **2000**, *112*, 5558.
- (51) Chipman, D. M. *J. Chem. Phys.* **1997**, *106*, 10194.
- (52) Zhan, C.-G.; Bentley, J.; Chipman, D. M. *J. Chem. Phys.* **1998**, *108*, 177.
- (53) Chipman, D. M. *J. Chem. Phys.* **2006**, *124*, 224111.
- (54) Vilkas, M. J.; Zhan, C.-G. *J. Chem. Phys.* **2008**, *129*, 194109.
- (55) Amovilli, C.; Filippi, C.; Floris, F. M. *J. Chem. Phys.* **2008**, *129*, 244106.
- (56) Cancès, E.; Mennucci, B. *J. Chem. Phys.* **2001**, *114*, 4744.
- (57) Chipman, D. M. *Theor. Chem. Acc.* **2002**, *107*, 80.
- (58) Mennucci, B.; Cammi, R.; Tomasi, J. *J. Chem. Phys.* **1998**, *109*, 2798.
- (59) Cossi, M.; Barone, V. *J. Phys. Chem. A* **2000**, *104*, 10614.
- (60) Zhan, C.-G.; Chipman, D. M. *J. Chem. Phys.* **1998**, *109*, 10543.
- (61) Zhan, C.-G.; Chipman, D. M. *J. Chem. Phys.* **1999**, *110*, 1611.
- (62) Floris, F.; Tomasi, J. *J. Comput. Chem.* **1989**, *10*, 616.
- (63) Floris, F. M.; Tomasi, J.; Pascual-Ahuir, J. L. *J. Comput. Chem.* **1991**, *12*, 784.
- (64) Rösch, N.; Zerner, M. C. *J. Phys. Chem.* **1994**, *98*, 5817.
- (65) Olivares del Valle, F. J.; Aguilar, M. A. *THEOCHEM* **1993**, *99*, 25.
- (66) Amovilli, C. *Chem. Phys. Lett.* **1994**, *229*, 244.
- (67) Amovilli, C.; Mennucci, B. *J. Phys. Chem. B* **1997**, *101*, 1051.
- (68) Schmidt, M. W.; Baldrige, K. K.; Boatz, J. A.; Elbert, S. T.; Gordon, M. S.; Jensen, J. H.; Koseki, S.; Matsunaga, N.; Nguyen, K. A.; Su, S. J.; Windus, T. L.; Dupuis, M.; Montgomery, J. A. *J. Comput. Chem.* **1993**, *14*, 1347.
- (69) Hermann, R. B. *J. Phys. Chem.* **1972**, *76*, 2754.
- (70) Smith, R.; Tanford, C. *Proc. Natl. Acad. Sci. U.S.A.* **1973**, *70*, 289.
- (71) Chothia, C. *Nature* **1974**, *248*, 338.
- (72) Reynolds, J. A.; Gilbert, D. B.; Tanford, C. *Proc. Natl. Acad. Sci. U.S.A.* **1974**, *71*, 2925.
- (73) Pierotti, R. A. *Chem. Rev.* **1976**, *76*, 717.
- (74) Sharp, K. A.; Nicholls, A.; Fine, R. F.; Honig, B. *Science* **1991**, *252*, 106.
- (75) Ashbaugh, H. S.; Kaler, E. W.; Paulaitis, M. E. *J. Am. Chem. Soc.* **1999**, *121*, 9243.
- (76) Huang, D. M.; Chandler, D. *J. Phys. Chem. B* **2002**, *106*, 2047.
- (77) Chandler, D. *Nature* **2005**, *437*, 640.
- (78) Floris, F. M.; Selmi, M.; Tani, A.; Tomasi, J. *J. Chem. Phys.* **1997**, *107*, 6353.
- (79) Amovilli, C.; Floris, F. M. *Phys. Chem. Chem. Phys.* **2003**, *5*, 363.
- (80) Ashbaugh, H. S.; Pratt, L. R. *Rev. Mod. Phys.* **2006**, *78*, 159.
- (81) Carlson, H. A.; Jorgensen, W. L. *J. Phys. Chem.* **1995**, *99*, 10667.

Multiconfigurational Second-Order Perturbation Theory Restricted Active Space (RASPT2) Studies on Mononuclear First-Row Transition-Metal Systems

Steven Vancoillie, Hailiang Zhao, Van Tan Tran, Marc F. A. Hendrickx, and Kristine Pierloot*

Department of Chemistry, University of Leuven, Celestijnenlaan 200F, B-3001 Heverlee, Belgium

S Supporting Information

ABSTRACT: A series of model transition-metal complexes, CrF_6 , ferrocene, $\text{Cr}(\text{CO})_6$, ferrous porphyrin, cobalt corrole, and FeO/FeO^- , have been studied using second-order perturbation theory based on a restricted active space self-consistent field reference wave function (RASPT2). Several important properties (structures, relative energies of different structural minima, binding energies, spin state energetics, and electronic excitation energies) were investigated. A systematic investigation was performed on the effect of: (a) the size and composition of the global RAS space, (b) different (RAS1/RAS2/RAS3) subpartitions of the global RAS space, and (c) different excitation levels (out of RAS1/into RAS3) within the RAS space. Calculations with active spaces, including up to 35 orbitals, are presented. The results obtained with smaller active spaces (up to 16 orbitals) were compared to previous and current results obtained with a complete active space self-consistent field reference wave function (CASPT2). Highly accurate RASPT2 results were obtained for the heterolytic binding energy of ferrocene and for the electronic spectrum of $\text{Cr}(\text{CO})_6$, with errors within chemical accuracy. For ferrous porphyrin the intermediate spin $^3\text{A}_{2g}$ ground state is (for the first time with a wave function-based method) correctly predicted, while its high magnetic moment ($4.4\mu_B$) is attributed to spin-orbit coupling with very close-lying $^5\text{A}_{1g}$ and $^3\text{E}_g$ states. The toughest case met in this work is cobalt corrole, for which we studied the relative energy of several low-lying $\text{Co}(\text{II})$ -corrole π radical states with respect to the $\text{Co}(\text{III})$ ground state. Very large RAS spaces (25–33 orbitals) are required for this system, making compromises on the size of RAS2 and/or the excitation level unavoidable, thus increasing the uncertainty of the RASPT2 results by 0.1–0.2 eV. Still, also for this system, the RASPT2 method is shown to provide distinct improvements over CASPT2, by overcoming the strict limitations in the size of the active space inherent to the latter method.

1. INTRODUCTION

Multiconfigurational perturbation theory based on a complete active space reference wave function (CASSCF/CASPT2)^{1,2} is a well-established computational quantum chemistry method for the study of relatively large transition-metal (TM) complexes, for which it is often seen as a preeminent wave function-based 'alternative' to density functional theory (DFT).³ The recent possibility of using multiconfigurational second-order perturbation theory based on a restricted active space reference wave function (RASSCF/RASPT2), instead of a complete active space reference wave function, has significantly reduced the computational cost associated with large active spaces.⁴ This has already led to a number of publications either to test how this new method can be used to achieve similar accuracy, while reducing the calculation time,^{5–7} and to overcome the size limitations of the active space in case of multicenter TM complexes,⁴ oligomeric hydrocarbons,⁸ and TMs bonded to extended π systems.^{9,10}

In a recent paper of Sauri et al.,⁶ several computational strategies were introduced for selecting and dividing the RASSCF/RASPT2 active space. This is more difficult than for the CASSCF/CASPT2 method because of the number of possible divisions of the active space that can be combined with different levels of excitations. Instead of a complete active space (CAS) where all possible distributions of the active electrons in the active orbitals are taken into account, a restricted active space

(RAS) is employed where the active orbitals are further subdivided into three subspaces: RAS1, RAS2, and RAS3. The orbitals in RAS1 are doubly occupied apart from allowing a maximum number of holes, while in RAS3 the orbitals are empty apart from allowing a limited number of electrons. In RAS2, all remaining active electrons are distributed over the orbitals, like in the original CAS. This whole procedure makes the RASSCF/RASPT2 method much less systematic than CASSCF/CASPT2. The previous RASPT2 study⁶ mainly focused on the electronic spectroscopy of organic molecules. Only two simple TM systems, CuCl_4^{2-} and the Ni atom, were considered, giving particular attention to the description of the so-called 3d double-shell effect¹¹ by means of RASPT2 rather than CASPT2.

The present paper aims at providing a more systematic overview of the possibilities and limitations of the RASSCF/RASPT2 method in the field of (first-row) TM chemistry. The specific electronic structure of TM complexes, originating from a partially filled d shell surrounded by a number of potential electron-donating or -accepting ligands, gives rise to specific strong correlation effects which cannot be handled by second-order perturbation theory and should therefore be included already in the reference wave function (be it of CAS or RAS type). Two types of such strong correlation effects are by now

Received: August 24, 2011

Published: October 20, 2011

generally known and have become the basis for the ‘standard’ rules for constructing the active space for TM systems.^{12–16}

- Nondynamic correlation effects connected to covalent TM–ligand interactions, giving rise to a set of bonding and antibonding molecular orbitals with mixed 3d–ligand character should be dealt with by including both combinations in the active space. Shortly, all valence molecular orbitals containing a significant metal d contribution should be active.
- The 3d double-shell effect. This correlation effect is important in systems with a high 3d occupation number, in particular if the d-orbital occupation changes between two states considered. It has to be dealt with in the reference wave function by including a second d shell, called either 3d’ or 4d, in the active space.

These two simple rules, however, are only the basis of an often much more complicated process of choosing active orbitals. For example, the first rule only suffices to describe the ground-state and metal-centered (MC, also called ligand field LF) excited states in complexes with weakly covalent bonds (for a counterexample, see the case of CrF₆ in Section 3.1). Moreover, it should always be kept in mind that the active space should not only include molecular orbitals involved in strong correlation effects but also orbitals that are or become partially occupied in the process studied. When describing, e.g., an electronic spectrum, the number of such orbitals may become quite large. The latter point is of course not specific to TM chemistry. However, in TM complexes the combination of covalency effects, double-shell effect, and the possible occurrence of different types of excited states [MC, charge-transfer (CT), either metal-to-ligand (MLCT) or ligand-to-metal (LMCT), and ligand-centered (LC)] may quite easily lead to a complex exercise of selecting active orbitals in a way that ensures a reasonable accuracy from the perturbation treatment, while keeping the size of the active space computationally manageable (i.e., up to about 16 orbitals in a CAS). The recent implementation of the RASPT2 method may have simplified this aspect of the calculational setup, as it allows for a less strict selection of the total number of active orbitals. However, the selection of the different subspaces and excitation levels (out of RAS1, into RAS3) within the RAS active space now becomes an additional crucial task that, in the case of TM complexes, has so far been virtually unexplored. How far can one go in moving orbitals involved in strong correlation effects, either due to covalent TM–ligand bonding or double-shell effects, from RAS2 into either RAS1 or RAS3, and which excitation level would then be required to maintain a similar accuracy in RASPT2 as compared to CASPT2? Should the orbitals that are involved in electronic excitations be included in RAS2 or not? Questions like this will be given an appropriate answer in the discussion below.

The aim of this contribution is to design reliable strategies for general purpose RASPT2 calculations in TM chemistry. For that purpose, we have included a series of TM systems and chemical problems that should cover a broad range of typical correlation problems encountered when computing TM complexes. The systems chosen are chromium hexafluoride CrF₆, chromium hexacarbonyl Cr(CO)₆, ferrocene Fe(Cp)₂ (Cp = cyclopentadienyl), iron oxide FeO and its anion FeO[−], ferrous porphyrin Fe(P) (P = porphyrin), and cobalt corrole Co(C) (C = corrole). With exception of the last system, all of these molecules have been studied before by means of CASPT2, although with

varying success.^{14,17–23} The specific problems encountered for each molecule will be introduced in the appropriate section, and a detailed investigation of possible solutions coming from RASSCF/RASPT2 will be presented.

2. COMPUTATIONAL DETAILS

All CASSCF/CASPT2 and RASSCF/RASPT2 calculations were performed with the MOLCAS-7 program²⁴ making use of relativistic atomic natural orbital (ANO-rcc) type basis sets.^{25,26} Unless noted otherwise, the contractions used are [7s6p5d3f2g1h] for the metal, [4s3p2d1f] on first-row atoms C, N, O, F, and [3s1p] on H. Scalar relativistic effects were included using a standard second-order Douglas–Kroll–Hess (DKH) Hamiltonian.^{27–30} In all CASPT2/RASPT2 calculations, the core electrons, i.e., 1s from (C, N, O, F) and 1s–2p from the metal, were kept frozen. All CASPT2 and RASPT2 calculations in this work were performed using the standard ionization potential electron affinity (IPEA) Hamiltonian³¹ (a zeroth-order Hamiltonian containing an IPEA shift = 0.25 au has become standard since MOLCAS-6.4). To avoid weak intruder states and improve convergence of the perturbational treatment, an imaginary level shift of 0.1 au was used.³² Furthermore, in all calculations advantageous use (in terms of computational times and disk storage needs) was made of the Cholesky decomposition of the electron repulsion integral matrix^{33–35} (with a threshold of $\delta = 10^{-6}$ au).

The specific choice of the active spaces used to construct the CASSCF and RASSCF reference wave function is of course different for each of the considered molecule and will therefore be presented in the appropriate section of the results. However, it is important to describe here the notation employed to label the CASSCF and RASSCF active spaces. For the CASSCF calculations, the traditional notation will be used, i.e., CAS(*n,a*), where *n* is the number of electrons included in the active space and *a* is the number of active orbitals. For the RASSCF calculations, the short notation RAS(*n,a*) will be maintained to denote the global active space. However, the RAS subpartitions and excitation levels are now specified by a longer notation, RAS(*n,l,m;i,j,k*) with *l* the maximum number of holes allowed in RAS1 and *m* the maximum number of electrons to enter RAS3. Active orbitals are labeled *i,j,k* (with *i + j + k = a*) and refer to those placed in RAS1, RAS2 and RAS3, respectively. Sometimes we will also use SD, SDT, SDTQ, etc., to denote the maximum excitation level out of RAS1 and/or into RAS3 (SD corresponding to *l = m = 2*, SDT to *l = m = 3*, ...).

In several cases, RASSCF calculations involving high excitation levels have been performed without performing an actual orbital optimization, i.e., they are in fact CI calculations based on orbitals obtained from a calculation with the same RAS but at a lower excitation level, usually SD. The primary motivation for considering this option is that RASSCF orbital optimizations, in the present version of the MOLCAS code, are really poorly convergent. Of course, such a procedure can only be successful if the effect of the higher excitations on the shape of the molecular orbitals is minimal. Extensive testing to prove this has been performed in case of CrF₆ (Section 3.1) and Co(C) (Section 3.5). Further specifications on the orbital optimization levels used will be given in each section.

Most calculations performed in this work are single-point calculations. Only for FeO/FeO[−], structures were optimized at the CASPT2/RASPT2 level. In all other calculations, structures were used that were taken either from experiment, from DFT

Table 1. Relative Energy (kcal/mol) of the Trigonal Prismatic and Octahedral Structure of CrF₆ at Different Computational Levels

orbitals	CASPT2(10,10)	RASPT2(42, <i>l,m</i> ;21,0,5)			RASPT2(42, <i>l,0</i> ;16,10,0)	
		SD	SDTQ	SDTQ56	SD	SDTQ
fully optimized	49.0	8.1	15.0		12.8	
taken from SD		8.1	15.4	16.5	12.8	17.8

structure optimizations, or from previously reported calculations. More precise details will be given in the appropriate sections.

3. RESULTS AND DISCUSSION

3.1. Relative Energies of the Octahedral and Trigonal Prismatic Structure of CrF₆. Considering correlation effects, the CrF₆ molecule does not exactly behave as a ‘typical’ TM complex. The extremely high formal oxidation state of chromium, +VI, gives rise to very strong covalency in the Cr–F bonding and correspondingly strong static correlation effects. Within octahedral symmetry, covalent-bond formation at the molecular orbital level is only possible within the representations e_g, t_{2g}. This type of covalency may adequately be described by a (10,10) active space, including the bonding and antibonding Cr 3d–F 2p combinations within these two representations. An early CASPT2 calculation along this line was performed to calculate the relative energies of the octahedral (O_h) and trigonal prismatic (D_{3h}) structure of CrF₆.¹⁷ This calculation, however, turned out to strongly overestimate the stability of the octahedral with respect to the trigonal structure, predicting an energy difference of 49.9 kcal/mol, while corresponding values of 14.4 and 16.9 kcal/mol were obtained from coupled-cluster singles and doubles and perturbative triple excitations [CCSD(T)] and DFT(B3LYP), respectively.^{36,37} The origin of the CASPT2 error was analyzed later¹⁴ by means of a RASSCF calculation, including all F 2p orbitals in the active space, and was found to be due to the occurrence of large contributions in the ground-state wave function of excitations from the nonbonding F 2p orbitals into Cr 3d. A more accurate perturbational treatment would therefore have to start from a reference wave function including all such excitations, being built from an active space including 36 electrons in 23 (Cr 3d + F 2p) orbitals. Such a calculation is out of reach of CASPT2 but can quite easily be performed with RASPT2. In fact, the RASPT2 calculations presented here were performed with an even larger RAS(42,26) space, including also the Cr 3p shell, which was found to mix quite strongly with some of the F 2p orbitals. Two subdivisions of this global active space were considered: (A) RAS(42,*l,m*;21,0,5) (*l* = *m* = 2,4,6), i.e., with RAS2 empty, F 2p and Cr 3p in RAS1, and Cr 3d in RAS3; (B) (42,*l,0*;16,10,0) (*l* = *m* = 2,4) with RAS3 empty, and RAS2 now consisting of the bonding and antibonding Cr 3d–F 2p combinations providing the CAS(10,10) space in the original CASPT2 study.¹⁷ The SDTQ calculations with active space A were performed either using fully optimized orbitals or using fixed orbitals taken from the corresponding SD calculation. For the SDTQ56 calculations with active space (A) and for SDTQ with active space B, only fixed orbitals were used. In order to compare the RASPT2 results with our previous CASPT2 calculations,¹⁷ the present CASPT2 and RASPT2 calculations

on CrF₆ were performed on O_h and D_{3h} structures taken from that work. The results are presented in Table 1.

The main point to be noted from Table 1 undoubtedly is the huge difference between the results obtained from either CASPT2 or RASPT2. Comparing CASPT2(10,10) with RASPT2-(42,2,0;16,10,0) we find that moving all F 2p–Cr 3d (and Cr 3p–3d) correlation effects from the perturbational treatment to the active space and treating them variationally with SD gives a lowering of the energy barrier by as much as 36 kcal/mol. The differences between the distinct RASPT2 treatments are much more limited. With RAS2 empty, the results obtained with only up-to-double RAS1 → RAS3 excitations are clearly too low. Up to sextuple excitations are necessary to provide a (more or less) converged RASPT2 result. Note that the two results obtained with SDTQ, making use of either orbitals obtained from the SD calculation or reoptimized, differ by less than 1 kcal/mol, thus indicating that orbital optimization at this level is not strictly necessary. After moving the bonding and antibonding Cr 3d–F 2p combinations to RAS2, significantly better results are obtained already at the SD level. Notably though, also in this case a further improvement by as much as 5 kcal/mol may be obtained by raising the excitation level to SDTQ. Unfortunately, the combination of a 10 orbital RAS2 with up to sextuple excitations is computationally unfeasible, such that it cannot be verified that the (42,4,0;16,10,0) is indeed (close to) converged. In any case, the two ‘best’ results obtained from RASPT2, 16.5 vs 17.8 kcal/mol, are rather close and also compare well to the earlier results obtained from DFT and CCSD(T), 14.2–16.9 kcal/mol.^{36,37} In conclusion, these calculations on CrF₆ present a clear-cut (first) example of a multiconfigurational problem that was unsolvable before with CASPT2 because of limitations in the size of the CAS space and for which the extension to RAS provides a considerable and very valuable improvement of the multiconfigurational perturbation treatment.

3.2. Heterolytic Binding Energy of Ferrocene. One of the first CASPT2 studies in organometallic chemistry, by Pierloot, Persson, and Roos (PPR) in 1995,¹⁹ concerned the equilibrium structure and binding energy of ferrocene. In particular, the heterolytic dissociation energy of the reaction Fe(Cp)₂ → Fe²⁺(⁵D) + 2Cp[−] was an important test case for this method. An accurate description of this property can only be obtained when all important (nondynamic) correlation effects connected to the covalent bonding between the Fe²⁺ ion and the two cyclopentadienyl (Cp[−]) ligands are adequately treated. Starting from an MP2 treatment, which was found to overestimate the binding energy by about 50 kcal/mol, a successful CASPT2 treatment was obtained by extending the reference wave function from Hartree–Fock to CASSCF with a (10,10) active space (specifications are given below). After correcting for basis set superposition errors (BSSE), the ‘best’ CASPT2 value for the heterolytic dissociation energy was 628 kcal/mol, in close

Table 2. ANO-*rcc* Basis Set Contractions Used for the CASPT2/RASPT2 Calculations on Ferrocene

atom	VTZP (I)	VQZP (II)	V8ZP/V5ZP (III)	V8ZP/V7ZP/V6ZP (IV)
Fe	6s5p3d2f1g	7s6p4d3f2g1h	10s9p8d6f4g2h	10s9p8d6f4g2h
C	4s3p2d1f	5s4p3d2f1g	6s5p3d2f1g	8s7p4d3f2g
H	3s2p1d	4s3p2d1f	5s3p2d1f	6s4p3d1f

Table 3. Basis set superposition errors (kcal/mol) obtained with CASPT2 and different basis sets

	basis I	basis II	basis III	basis IV
Without Fe (3s,3p) Correlation				
Fe ²⁺ + 2 Cp(ghost)	4.8	2.4	0.1	0.1
Cp ⁻ + Fe(ghost) + Cp(ghost)	9.7	3.7	3.1	1.1
sum	24.2	9.8	6.3	2.3
Including Fe (3s,3p) Correlation				
Fe ²⁺ + 2 Cp(ghost)	30.9	21.7	0.6	0.9
Cp ⁻ + Fe(ghost) + Cp(ghost)	9.7	3.7	3.1	1.1
sum	50.3	29.1	6.8	3.1

agreement with the value of (635 ± 6) kcal/mol reported from experiment³⁸ for the heterolytic bond disruption enthalpy of ferrocene. Admittedly, the close agreement with experiment was to some extent fortuitous, being the result of a cancellation of errors related to the (limited) size of the basis sets, the absence of zero-point vibrational energy (ZPVE) and thermal corrections and the rather crude treatments of relativistic effects and of Fe (3s,3p) (semicore) correlation. As a matter of fact, the (ANO-s) basis set used for iron in the PPR study¹⁹ is not properly designed for treating (3s,3p) correlation,³⁹ giving rise to huge BSSE on Fe. Therefore, semicore correlation was in the PPR study only included for the calculation of the ¹I–⁵D excitation energy in Fe²⁺ but was omitted for the dissociation step. A critical analysis of the different sources of errors in the CASPT2 treatment of ferrocene was presented in a later paper by Klopper and Lüthi,⁴⁰ who concluded that, when extrapolated to the basis set limit, the heterolytic bond disruption enthalpy of ferrocene obtained from CASPT2(10,10) would be Δ*H*₂₉₈⁰ = 657 kcal/mol, around 20 kcal/mol higher than the experimental value of (635 ± 6) kcal/mol. This would suggest that CASPT2, based on a CAS(10,10) reference wave function, inherently quite strongly overestimates the binding energy in this molecule. (Notably though, extrapolated to the basis set limit, the CASPT2 binding energy was very similar to the result obtained using coupled cluster theory.)⁴¹

In this work, the heterolytic binding energy of ferrocene is revisited by means of CASPT2 and RASPT2. Developments in the MOLCAS software over the last years have made it possible to strongly reduce the errors afflicting the reliability of the early CASPT2 results. In particular new ANO-*rcc* basis sets, constructed making use of the Douglas–Kroll Hamiltonian, should provide an accurate description of scalar relativistic effects.^{25,26} For the TM, these basis sets were designed to treat semicore correlation, leading to more accurate description of this type of correlation and smaller BSSE. Moreover, the latter errors may be circumvented by making use of much more extended contracted basis sets, thanks to the recent advent in MOLCAS of the Cholesky decomposition technique to approximate the two-electron integrals and its combination with CASPT2/RASPT2.^{33–35} Four different combinations of ANO-*rcc*

contraction schemes, denoted I–IV and shown in Table 2, were used in this section, the largest combination giving rise to 1457 contracted functions.

3.2.1. Computational Details. Ferrocene has a ground-state geometry where the two pentadienyl rings are eclipsed, with the staggered conformation approximately 1 kcal/mol higher in energy. The point group symmetry of the eclipsed structure is *D*_{5h} and the calculations were performed within the highest abelian subgroup symmetry *C*_{2v}.

Single-point calculations on the experimental geometry of ferrocene⁴² were performed, with the following distances: *R*(C–C) = 1.440 Å, *R*(C–H) = 1.104 Å, and *R*(Fe–Cp) = 1.66 Å. For free Cp⁻ a *D*_{5h} structure was optimized by means of MP2 making use of the smallest VTZP basis set (I), giving *R*(C–C) = 1.397 Å and *R*(C–H) = 1.079 Å. All binding energies reported include a correction for the BSSE, obtained by means of the full counterpoise method, i.e., from a calculation either on Fe²⁺ adding ghost orbitals of the two rings and on Cp⁻ adding ghost orbitals of Fe²⁺ plus the other ring. BSSE were calculated for all combinations of methods and basis sets. However, using the same basis sets the differences between the RASPT2 and CASPT2 results are insignificantly small (<1 kcal/mol). Therefore, only the CASPT2 results are given in Table 3. To be able to compare the calculated results for the binding energies to the experimental value of the heterolytic bond disruption enthalpy, Δ*H*₂₉₈⁰ = (635 ± 6) kcal/mol,³⁸ a correction for the difference in zero-point vibrational energy (–9.3 kcal/mol), and the thermal correction to the reaction enthalpy (13/2 *RT* + the difference in vibrational energies = 2.9 kcal/mol) were estimated from DFT optimizations and frequency analyses on ferrocene and Cp⁻, making use of the BP86 functional and extended basis sets (def2-QZVPP on Fe, def2-TZVP for C,H). Rather than adding these corrections to all calculated CASPT2/RASPT2 binding energies, we have chosen to compare the uncorrected data to a ‘corrected’ experimental heterolytic binding energy of (641 ± 6) kcal/mol.

3.2.2. CASPT2 Calculations: Results and Discussion. First, a set of CASPT2 calculations was performed with different basis sets, making use of the same CAS(10,10) active space that was used in the PPR study.¹⁹ This active space includes those pairs of orbitals that are directly involved in covalent Fe–Cp interactions: the (4,5)*e*₁^g shells, consisting of the bonding and antibonding combinations of Cpπ and the (formally empty) Fe (*d*_{xz},*d*_{yz}) orbitals, and the (4,5)*e*₂^g shells, incorporating the backdonation from the (formally doubly occupied) Fe (*d*_{xy},*d*_{x²–y²}) orbitals into the Cpπ* orbitals of the same symmetry. We note though that the *S**e*₂^g shell also contains significant Fe 3*d*^g character (for a more detailed analysis of the covalent interactions in ferrocene see ref 19). On top of this (8,8) space, the essentially nonbonding Fe 3*d*_{z²} orbital of *a*₁^g symmetry and its correlating 3*d*^g orbital were also included. In the heterolytic dissociation limit, this active space is split into a CAS(6,8) space on Fe²⁺ (five 3*d* and three correlating 3*d*^g orbitals) and a CAS(4,2) space on Cp⁻, the latter therefore being described by MP2 rather than CASPT2. However, since the dissociation proceeds to the Fe(II) ⁵D ground

Table 4. Binding Energies^a (E_b) for the Reaction $\text{FeCp}_2 \rightarrow \text{Fe}^{2+}({}^5\text{D}) + 2\text{Cp}^-$ (kcal/mol)

	basis I	basis II	basis III	basis IV
Without Fe (3s,3p) Correlation				
CASPT2(10,10)	618.5	635.3	639.1	641.1
RASPT2(14,2,2;3,8,7)	605.1	622.4	625.5	627.1
Including Fe (3s,3p) Correlation				
CASPT2(10,10)	633.2	650.9	655.1	657.2
RASPT2(14,2,2;3,8,7)	619.1	637.2	640.5	641.4
RASPT2(14,4,4;3,8,7)	619.2			
RASPT2(14,6,6;3,8,7)	618.8			
RASPT2(14,2,2;7,0,11)	610.5	628.6	632.4	
RASPT2(14,4,4;7,0,11)	630.0	647.7	651.4	
RASPT2(14,6,6;7,0,11)	632.0			
RASPT2(14,8,8;7,0,11)	632.2			
expt		(641 ± 6)		

^a Corrected for BSSE.

state with all 3d orbitals at least singly occupied, we chose to calculate the energy of the Fe^{2+} fragment using a full 3d' shell in the active space of the fragment iron cation, i.e., CAS(6,10) rather than with the unbalanced CAS(6,8) space.

All calculated binding energies are presented in Table 4. At the CASPT2 level, calculations with and without Fe (3s,3p) correlation were performed, and the results in Table 4 clearly indicate that iron semicore correlation is important and should be included when calculating the binding energy, increasing the results with up to 16 kcal/mol. However, as indicated by Table 3, including the Fe(3s,3p) electrons in the correlation treatment drastically increases the BSSE on iron, at least with the smallest contractions (basis I, II). As the BSSE is very large, its (counterpoise) correction may not be very reliable. Therefore, we decided to extend the basis sets up to a point where the BSSE would become small enough, i.e., lower than the errors that might arise from the CASPT2/RASPT2 treatment. Comparing the results with different basis sets in Table 4, one can observe an important increase of the binding energy, about 22–24 kcal/mol, with the size of the basis sets. Making use of the largest basis set IV and including iron semicore correlation, the heterolytic binding energy of ferrocene obtained from CASPT2 amounts to 657.2 kcal/mol. This result is considerably larger than the binding energy from the original CASPT2 study (628 kcal/mol).¹⁹ As compared to experiment, CASPT2 seems to overestimate the total strength of the Fe–Cp bonds by at least 10 kcal/mol. Given the extended size of the basis sets employed here, the source of the remaining error should be traced back to the CASPT2 method itself and more specifically to the limited size of the CAS(10,10) reference wave function.

3.2.3. RASPT2 Calculations: Results and Discussion. In order to investigate the effect of the active space on the binding energy, a second series of calculations was performed with a larger active space, however now making use of RASPT2 instead of CASPT2. The CAS(10,10) space was extended with the $6e_1(\text{Cp}\pi)$, $3e_2(\text{Cp}\pi^*)$ couple as well with the remaining four Fe 3d' orbitals, thus giving RAS(14,18). As for CrF_6 in the previous section, two options for subdividing this active space were investigated, i.e., (A) with RAS2 comprising the $(4,5)e_1''$, $(4,5)e_2''$ shells involved in covalent Fe 3d–Cp π interactions \rightarrow RAS(14, l,m ;3,8,7), and (B)

with an empty RAS2 space \rightarrow RAS(14, l,m ;7,0,11). Upon dissociation, the global RAS(14,18) active space is split into a RAS(6,10) space on iron (all 3d, 3d' orbitals) and a RAS(4,4) π space on Cp^- (while taking care of the appropriate subdivisions corresponding to (A) or (B) in FeCp_2). For the fragments, an excitation level $l = m = 2$ was used in all calculations. In order to obtain a fully size-extensive treatment, this means that up to sextuple excitations should be allowed for FeCp_2 . In order to check this, test calculations with $l = m = 4,6$ were performed with the smallest basis sets (making use of full CI calculations with the orbitals taken from the SD calculation).

We first consider the RASPT2 calculations performed with the RAS(14,2,2;3,8,7) space. Such calculations were performed with all four basis sets and both with and without semicore correlation. As can be seen from Table 4, the effect of Fe (3s,3p) correlation on the binding energy is similar for the RASPT2 and CASPT2 calculations, giving rise to a bond strengthening by about 15 kcal/mol. Both with and without (3s,3p) correlation, the RASPT2-(14,2,2;3,8,7) binding energy is systematically lower than the CASPT2 binding energy by 13 kcal/mol, a value which does not significantly depend on the size of the basis set. We also note the very small differences between the results obtained at different excitation levels $l, m = 2-6$. When keeping in RAS2 the eight orbitals that are responsible for the most important nondynamic correlation effects, $(4,5)e_1''$, $(4,5)e_2''$, the RASPT2 treatment of FeCp_2 converges already at the SD level. However, it is clear that the variational treatment in the RASSCF wave function of extra important correlation, i.e., the full Fe 3d double-shell effect and the most important $\text{Cp}^- \pi-\pi^*$ correlation effects, has a substantial effect on the result of the perturbational treatment, lowering the binding energy by more than 10 kcal/mol. With the largest basis set IV and including semicore correlation, the heterolytic dissociation energy of FeCp_2 obtained from RASPT2 amounts to 641.1 kcal/mol, as compared to the experimental value of (641 ± 6) kcal/mol.

Less satisfactory results are obtained from the RASPT2 calculations with an empty RAS2 space. As Table 4 shows, the binding energies obtained from the RASPT2(14,2,2;7,0,11) calculations are considerably lower, by 8–9 kcal/mol, than the corresponding results with RASPT2(14,2,2;3,8,7). On the other hand, between SD and SDTQ the binding energies from the empty RAS2 calculations are increased by 19–22 kcal/mol, while the calculations with the eight-orbital RAS2 are already converged at the SD excitation level. With an empty RAS2, convergence is reached only with SDTQ rather than with SD. However, the ultimate binding energy, 651.4 kcal/mol (with basis III, a slightly higher value may be anticipated for basis IV) is now higher by 11 kcal/mol than the result obtained from the eight-orbital RAS2 calculation.

The question may of course be raised as to why the two sets of RASPT2 calculations presented in Table 4 converge to such strongly different values of the binding energy. Should they not both converge to the same full CI limit, being then also the answer that would be obtained from a full CASPT2(14,18) calculation? That this is not the case is related to the way a RASPT2 calculation is designed in its current implementation in the MOLCAS code. As with CASPT2, a RASPT2 calculation starts with the construction of an effective Fock matrix. In the CASPT2 case, this Fock matrix is diagonalized within the three subspaces inactive/inactive, active/active, and external/external. Note that unitary transformations of the orbitals within these three subspaces do not affect the overall CASSCF wave function.

Table 5. Relative Energies of the Most Important LF and CT Excited States in Cr(CO)₆

method	¹ T _{1g}	¹ T _{2g}	a ¹ T _{2u}	a ¹ T _{1u}	b ¹ T _{2u}	b ¹ T _{1u}
CASPT2(10,16)	4.92	5.35	3.88	4.31(0.20)	4.68	5.16(2.44)
MS-CASPT2(10,16)	4.92	5.35	3.87	4.29(0.42)	4.69	5.18(2.18)
RASPT2(10,3,3;5,0,11) ^a	4.50	4.94	3.47	3.89(0.22)	4.26	4.91(2.55)
RASPT2(10,3,3;5,0,11) ^b	4.65	5.09	3.62	4.03(0.22)	4.41	5.06(2.54)
RASPT2(10,5,5;5,0,11) ^c	4.86	5.28	3.69	4.11(0.22)	4.47	5.10(2.53)
RASPT2(10,5,5;5,0,11) ^d	4.88	5.31	3.71	4.14(0.22)	4.50	5.12(2.54)
RASPT2(10,7,7;5,0,11) ^e	4.92	5.35	3.74	4.16(0.22)	4.61	5.14(2.54)
RASPT2(10,0,2;0,10,6)	4.92	5.35	4.06	4.50(0.17)	4.89	5.43(2.37)
RASPT2(10,0,3;0,10,6)	4.92	5.35	3.87	4.30(0.20)	4.67	5.14(2.47)
RASPT2(10,0,4;0,10,6)	4.92	5.35	3.86	4.29(0.20)	4.67	5.14(2.43)
RASPT2(10,0,2;0,10,14)	4.75	5.08	3.96	4.40(0.20)	4.83	5.27(2.56)
RASPT2(10,0,3;0,10,14)	4.98	5.23	4.07	4.50(0.20)	4.91	5.43(2.61)
RASPT2(10,0,4;0,10,14)	4.98	5.23	4.07	4.50(0.20)	4.91	5.42(2.57)
MS-RASPT2(10,0,4;0,10,14)	4.98	5.23	4.06	4.47(0.46)	4.91	5.45(2.26)
expt ^f				4.44(0.25)		5.48(2.3)

^a Using SD for GS. ^b Using SDT for GS. ^c Using SDTQ for GS. ^d Using SDTQ5 for GS. ^e Using either SDTQ56 or SDTQ567 for GS. ^f From Beach and Gray.⁴³

Within the basis of the resulting ‘canonical’ orbitals, all off-diagonal Fock matrix elements connecting active orbitals are zero. This is, however, not the case in RASPT2. Here, orbital rotations that couple different RAS subspaces, i.e., RAS1 and RAS2, RAS1 and RAS3, and RAS2 and RAS3, are not allowed. Thus, the diagonalization of the active part of the Fock matrix is not complete in RASPT2, as it is in CASPT2, but instead nonzero elements may remain in the parts that couple different RAS subspaces. In the current implementation of RASPT2 these elements are ignored, and the coupling is thus not accounted for in the perturbation. If the coupling elements are large, then this may lead to RASPT2 results that are strongly different from the corresponding CASPT2 solution or from an alternative RASPT2 solution which does include the strongly interacting orbitals in the same RAS subspace. This is the case here, with the bonding–antibonding Fe 3d–Cpπ combinations allowed to rotate among each other in the calculations with the (4,5)e₁[′], (4,5)e₂[′] couples in RAS2, but not in the empty RAS2 calculations, the former calculation obviously giving superior RASPT2 results. The present example indicates that, in order to provide an accurate description of the metal–ligand binding in covalently bonded TM complexes by means of RASPT2, the orbitals involved in the covalent interaction should preferably be kept together in RAS2.

3.3. Electronic Absorption Spectrum of Cr(CO)₆. The UV gas phase absorption spectrum of Cr(CO)₆ has been intensively studied in the past, both experimentally and theoretically, and its assignment is by now well established. The experimental spectrum with its assignment was first reported in 1968.⁴³ Two intense bands at relative energies of 4.44 and 5.48 eV were assigned as symmetry allowed ¹A_{1g} → ¹T_{1u} MLCT transitions, whereas two weak features at the red tail of the spectrum were originally assigned as ligand-field (LF) excitations. The latter assignment was, however, revised in the first CASPT2 study of the spectrum,²⁰ which indicated that the two singlet LF states, ¹T_{1g} and ¹T_{2g}, should in fact be situated at much higher energies than originally proposed and that the absorption features at the red tail of the spectrum are rather due to orbitally forbidden MLCT excitations. The latter assignment was soon confirmed in

a ΔSCF-DFT study by Pollak et al.,⁴⁴ who also proposed a reassessment of the role of the LF states in the photochemistry of Cr(CO)₆ and other metal–carbonyl complexes. Starting from then, the electronic spectra of Cr(CO)₆ and other metal carbonyl complexes M(CO)₆ (M = Cr, Mn⁺, V⁺, Mo, W) have frequently been used in benchmarking studies by means of time-dependent DFT (with different functionals),^{45,46} DFT/multireference configuration interaction (DFT/MRCI),⁴⁷ CASPT2,⁴⁸ and coupled-cluster methods.^{48,49}

Cr(CO)₆ is an octahedral molecule with a closed-shell ¹A_{1g} ground state corresponding to the configuration 2t_{2g}⁶6e_g⁰. The bonding in this molecule is built from two interactions with quite strong covalent character: σ-donation from CO into the formally empty 6e_g (Cr 3d) orbitals, counteracted by π backdonation from the fully occupied 2t_{2g} (Cr 3d) orbitals into the COπ* orbitals of the same symmetry. In order to describe the nondynamic correlation effects connected to these covalent interactions, a CAS(10,10) active space, including the Cr 3d orbitals and their bonding 5e_g and antibonding 3t_{2g} counterparts, is necessary. However, this CAS(10,10) space is not appropriate to describe the two most intense bands in the electronic spectrum, of ¹T_{1u} symmetry, corresponding to charge-transfer excitations into the 9t_{1u} and 2t_{2u} COπ* shells. To describe these states, the ‘basic’ CAS(10,10) space has to be extended with both these shells, giving a CAS(10,16) space. At the time of our original CASPT2 study of the Cr(CO)₆ spectrum, an active space of this size was computationally out of reach. This is no longer the case now, and in this work, we will start from a CASPT2(10,16) calculation of the most important bands in the Cr(CO)₆ spectrum. Starting from this CASPT2(10,16) calculation, several RAS spaces are then constructed, serving two goals: (A) to explore the adequacy of RASPT2 with the same global RAS(10,16) space but with different RAS partitions/excitation levels and (B) to extend the (10,16) active space with extra orbitals, finding ways to improve the accuracy of the calculations by comparing the results obtained for the excitation energies (oscillator strengths) of the two ¹T_{1u} states to the experimental data: 4.44 eV (0.25) and 5.48 eV (2.3). We will focus on just six excited states in the vertical spectrum of Cr(CO)₆, namely the two singlet LF states ¹T_{1g} and

${}^1T_{2g}$, both corresponding to a $2t_{2g} \rightarrow 6e_g$ excitation, and four MLCT states a,b^1T_{1u} and a,b^1T_{2u} corresponding to excitations out of $2t_{2g}$ into either $9t_{1u}$ or $2t_{2u}$. All calculations were performed within D_{2h} symmetry but with extra restrictions to prevent mixing of orbitals belonging to different representations within O_h . State-average calculations were performed for the two LF excited states ${}^1T_{(1,2)g}$ and for the four MLCT states $a, b^1T_{(1,2)u}$ since these sets of states belong to the same symmetry species within D_{2h} . The calculations were performed for the experimental octahedral structure of $\text{Cr}(\text{CO})_6$ with a Cr–C bond distance of 1.914 Å and a C–O distance of 1.140 Å.⁵⁰ The results are collected in Table 5.

The first two lines in Table 5 contain the CASPT2 data for the excitation energies and oscillator strengths obtained starting from the CAS(10,16) reference space. Both single-state and multistate (MS-CASPT2) results are given, the latter allowing for remixing of the wave functions of the pairs of ${}^1T_{1u}$ and ${}^1T_{2u}$ states under the influence of dynamical correlation (i.e., the CASPT2 treatment).⁵¹ One can see that the energetic effect of the MS algorithm is very limited, whereas its effect on the relative oscillator strength of both allowed (${}^1T_{1u}$) CT transitions is more significant. In fact, in absolute terms both pairs of oscillator strengths quite closely agree with experiment (the last line in Table 5). At the single-state level, the relative intensity of the second with respect to the first ${}^1T_{1u}$ state is slightly too high (12.2 versus 9.2 experimentally), whereas the MS treatment overcorrects, leading to a too small ratio (5.2). A detailed overview of the calculated excitation energies and corresponding spectral intensities at different computational levels for the two main bands in the UV absorption spectrum of $\text{Cr}(\text{CO})_6$ was reported very recently.⁴⁷ We note that the values of the oscillator strengths obtained from the present CASPT2(10,16) calculations are, together with the DFT/MRCI results from ref 47 (0.234 for a^1T_{1u} , 2.203 for b^1T_{1u}), superior to any of the previously calculated values for this property. The same is, however, not the case for the CASPT2(10,16) excitation energies. As Table 5 indicates, the calculated excitation energies are too low, with the largest error, 0.30–0.32 eV, for the b^1T_{1u} state. Even though the present and previous CASPT2 calculations show an accuracy which is good enough to provide a qualitative correct assignment of the $\text{Cr}(\text{CO})_6$ UV absorption spectrum, there is obvious room for improvement, coming from RASPT2 calculations with an extended active space.

Before moving to these calculations it is, however, necessary to test the accuracy that can be obtained with different RAS(10,16) subpartitions/excitation levels with respect to the full CASPT2(10,16) results. In Table 5 two different RAS(10,16) subpartitions have been considered. In the first case, RAS2 is empty, and the different states in the spectrum are described by RAS1 \rightarrow RAS3 excitations, starting from the closed-shell HF wave function of the ground state and allowing different maximum excitation levels. Since a single RAS1 \rightarrow RAS3 excitation is involved already in the primary configuration of all excited states, the excitation levels given in Table 5 are uneven, i.e., an even number (doubles, quadruples, ...) to describe correlation on top of the single excitation creating the open shell. When comparing the energy of the excited states to the energy of the closed-shell ground state, it is not obvious a priori whether the same or one lower excitation level should be used for the latter state. Results of both approaches have therefore been included in Table 5. Their difference reflects the effect of the additional excitation on the ground-state correlation treatment and disappears as the latter treatment converges.

As the results indicate, these RASPT2 calculations suffer from the same problems as noted already for the calculations with empty RAS2 for ferrocene, i.e., (a) slow convergence of the relative energies with the excitation level, and (b) the convergence limit (both for the excitation energies and oscillator strengths) does not equal the CASPT2 result, at least not for the CT excitations. The origin of these failures has been analyzed in detail in the previous section (Section 3.2). As for ferrocene, considerably better results may be obtained by including in RAS2 the 10 orbitals involved in the covalent Cr–CO interactions, leaving RAS1 empty and including only the t_{1u} and t_{2u} shells in RAS3. As can be seen from Table 5, the RASPT2(10,0; m ;0,10,6) results are converged already at the SD excitation level for the LF states (adding triples cannot affect the energy of these gerade states as all RAS2 \rightarrow RAS3 excitations are $g \rightarrow u$) and at the SDT level for the CT states, with excitation energies and oscillator strengths that are virtually equal to the corresponding CASPT2(10,16) results. As a general rule, we would therefore like to state again that, in order to obtain accurate RASPT2 results for relative energies in (first-row) TM systems, all ligand orbitals involved in covalent metal–ligand interactions should be kept together with the metal 3d orbitals in the RAS2 space.

Starting from the second set of RASPT2 calculations, the global active space can now straightforwardly be extended with additional RAS3 orbitals that might be expected to play an important role in the correlation treatment, i.e., (a) an extra 3d shell (t_{2g}, e_g) providing an improved description of the 3d double-shell effect, and (b) the remaining CO π^* orbitals of t_{1g} symmetry. RASPT2 results obtained with this RAS(10,0; m ;0,10,14) space are shown in the last block in Table 5. One can see that up-to-triple RAS2 \rightarrow RAS3 excitations are now necessary to reach convergence for both the LF and CT states. At the SDT or SDTQ level, the RASPT2 results obtained with this 24-orbital active space are, however, significantly different from the corresponding 16-orbital results. This is particularly true for the four CT states, which are blue-shifted by 0.2–0.3 eV. The calculated excitation energies of the two allowed ${}^1T_{1u}$ states now approach their experimental band positions to within 0.06 eV; after allowing remixing in a MS treatment, the difference is further decreased to 0.03 eV. The effect of extending the active space on the calculated oscillator strengths is hardly significant. On the whole, the present calculations clearly illustrate the added value of RASPT2 with respect to CASPT2 for the accurate description of the electronic spectra of organometallic systems.

3.4. Spin-State Energetics in Ferrous Porphyrin. Because of the important role played by iron porphyrins in biological processes (as the active centers or prosthetic groups of heme-proteins), the electronic structure of the Fe(II) ion in ferrous porphyrins has since long been the subject of experimental and theoretical investigations. The most important experimental data were collected already many years ago, between 1972 and 1985.^{52–60} On the other hand, computational chemists became interested in the problem since the beginning of the 1980s and by now a large number of computational studies have been reported.^{14,21,61–73} With six 3d electrons, the ground-state spin multiplicity of Fe(II) complexes may be either low-, intermediate-, or high-spin ($S = 0, 1, \text{ or } 2$, respectively). In Fe(II) porphyrins, all three spin states may occur, depending on the coordination number and the environment of the iron ion. Experimental data^{52–54} for the four-coordinate ferrous porphyrins tetraphenylporphyrinatoiron(II) (FeTPP)⁵² and octaethylporphyrinatoiron(II) (FeOEP)^{54,55} pointed to a triplet ground state

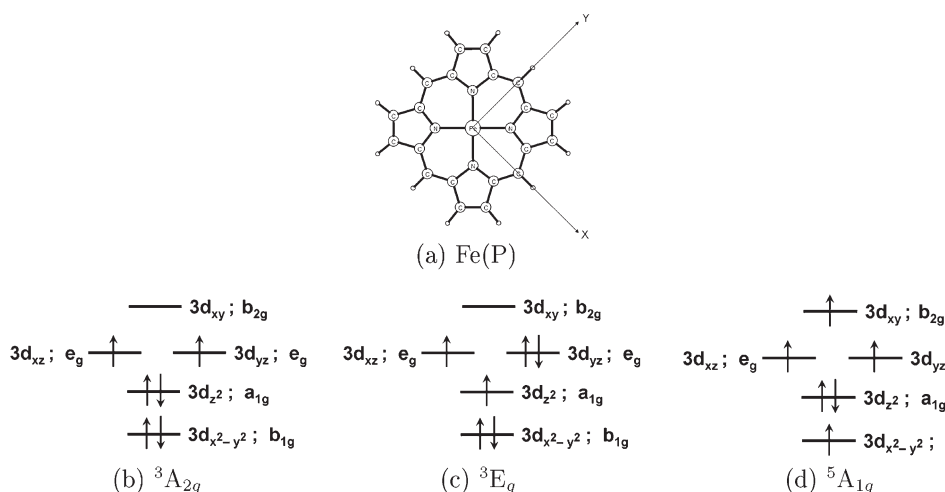


Figure 1. (a) Molecular structure of Fe(P), (b–d) orbital occupancy in the ${}^3A_{2g}$, ${}^3E_g(A)$, and ${}^5A_{1g}$ states.

but disagree on the details of the electronic configuration. For FeTPP, a ${}^3A_{2g}$ ground state was indicated by Mössbauer,^{52,56} magnetic,⁵⁷ or proton nuclear magnetic resonance measurements.^{53,58} On the other hand, resonance Raman spectra of FeOEP were originally interpreted in terms of a 3E_g ground state.⁵⁴ The magnetic moments measured for these two complexes, $4.4 \mu_B$ for FeTPP, and $4.6\text{--}4.7 \mu_B$ for FeOEP,^{55,59} are considerably higher than the pure triplet spin value $2.83 \mu_B$. This has been explained by the occurrence of one^{53,56,58} or two⁵⁷ low-lying triplet states (3E_g , ${}^3B_{2g}$) mixing through spin–orbit coupling (SOC) with the ${}^3A_{2g}$ ground state.

Obtaining accurate predictions of the spin state energetics in ferrous porphyrins has been shown in the past to be a real challenge for multiconfigurational perturbation theory (and other wave function methods). So far, all attempts to describe the relative energy of the low-lying states in free Fe(II) porphyrin (denoted as Fe(P) in what follows) have invariably predicted a high-spin ground state ${}^5A_{1g}$. A first attempt by means of multi-reference Møller–Plesset perturbation theory⁶⁵ found the ${}^5A_{1g}$ at 0.37 eV below the lowest triplet state, 3E_g . In a subsequent CASPT2 calculation, the lowest triplet state was found to be ${}^3A_{2g}$ with 3E_g at 0.02 eV. However, the ${}^5A_{1g}$ state was now even further stabilized with respect to the lowest triplet states, lying as much as 0.83 eV below ${}^3A_{2g}$. In 2003, it was shown by Pierloot¹⁴ that the CASPT2 description of the relative spin state energetics may be considerably improved by an adequate choice of the active space, i.e., by including next to the Fe 3d orbitals, a set of five 3d' orbitals to describe the 3d double-shell effect as well as one Fe–N σ bonding orbital to account for nondynamic correlation effects associated with the covalent-bonding character of the Fe 3d–N(P) σ bonds. With this CAS(8,11) space, the ${}^3A_{2g}$ – ${}^5A_{1g}$ splitting was reduced to 0.44 eV. In the same article, it was also shown that including eight porphyrin π orbitals (the four highest occupied and four lowest unoccupied) in the active space does not affect the spin state energetics to any significant extent. More recently, the unbalanced description of the intermediate- and high-spin states in Fe(P) could be partially traced back to the formulation of the zeroth-order Hamiltonian $\hat{H}^{(0)}$ in the original formulation of CASPT2, which would systematically favor high-spin over low-spin states. In 2004 the alternative IPEA modified $\hat{H}^{(0)}$ was proposed and implemented as the standard (IPEA shift of 0.25 au) in MOLCAS.³¹ Making use of the improved $\hat{H}^{(0)}$ and more

extended basis sets, still using the same CAS(8,11) space as before, the CASPT2 ${}^3A_{2g}$ – ${}^5A_{1g}$ splitting was further reduced to 0.19 eV (0.28 eV after correcting for the difference in zero-point vibrational energy).²¹ A similar error was also found for the splitting between the lowest triplet and quintet state in small heme model systems (for which CCSD(T) results could be used as a benchmark) and in the five-coordinated Fe(P)(Im) complex (Im = imidazole).²² From this it was concluded that the CASPT2 method has a systematic error of around 0.2 eV in favor of the high-spin state for ferrous heme complexes. Obviously, this error can only be explained by either the limited size of the active space or by limitations in the basis sets. Both options are explored in this work.

3.4.1. Computational Details. The basis sets used in the most recent CASPT2 calculations on Fe(P)²¹ already consisted of quite extended ANO-rcc type sets. However, they did not provide a fully balanced treatment of the porphyrin ring, in that a larger contracted [4s3p2d1f] set on N was combined with a smaller [4s3p1d] set on C. In this work, the contracted basis set on C is also enlarged to [4s3p2d1f]. Other than that, the basis sets used in this work (ANO-rcc contracted to [7s6p5d3f2g1h] for Fe, [4s3p2d1f] for N, C, and [3s1p] for H) are the same as in ref 21.

CASPT2/RASPT2 calculations were performed for 11 low-lying states in Fe(P), with varying occupations of the Fe 3d orbitals and with different spins ($S = 0\text{--}2$). The orientation of the molecule in the xy -plane as well as the symmetry (in D_{4h}) and occupations numbers of the Fe 3d orbitals in the three lowest-lying states are shown in Figure 1. All calculations were performed at structures obtained from DFT, making use of the PBE0 functional and extended basis sets (def2-QZVPP on Fe, and def2-TZVPP on other atoms) and employing the Turbo-mole code.^{74,75} Geometry optimizations were performed within D_{2h} symmetry, except for the ${}^3E_g(A)$ and 5E_g states, for which D_{2h} symmetry was used (pointing to a very weak Jahn–Teller distortion). Individual DFT structures could be obtained for all states reported in Table 6, except for the second ${}^3E_g(B)$ state, the 1E_g state, and the two open-shell singlet states ${}^1A_{1g}(A)$ and ${}^1A_{2g}$. The CASPT2/RASPT2 calculations for the ${}^3E_g(B)$ and 1E_g states were performed using the structure of the lower-lying ${}^3E_g(A)$ state, whereas for the open-shell singlets the structure of the ${}^3A_{2g}$ state (belonging to the same configuration) was used.

Table 6. Adiabatic Relative Energies (eV) of the Low-Lying Electronic States of Fe(P)

state	electronic configuration	CASPT2 (8,11)	RASPT2 (8,0,2;0,6,5)	RASPT2 (34,2,2;13,6,16)	CASPT2 (16,15)	RASPT2 (16,2,0;4,11,0)
$^3A_{2g}$	$(d_{x^2-y^2})^2(d_{z^2})^2(d_{xz},d_{yz})^2(d_{xy})^0$	0.00	0.00	0.00	0.00	0.00
$^5A_{1g}$	$(d_{x^2-y^2})^1(d_{z^2})^2(d_{xz},d_{yz})^2(d_{xy})^1$	-0.09	-0.19	-0.20	0.04	0.05
$^3E_g(A)$	$(d_{x^2-y^2})^2(d_{z^2})^1(d_{xz},d_{yz})^3(d_{xy})^0$	0.09	0.08	0.09	0.10	0.10
$^3E_g(B)$	$(d_{x^2-y^2})^1(d_{z^2})^2(d_{xz},d_{yz})^3(d_{xy})^0$	0.87	0.86	0.91	0.92	0.92
5E_g	$(d_{x^2-y^2})^1(d_{z^2})^1(d_{xz},d_{yz})^3(d_{xy})^1$	0.12	0.01	-0.01	0.21	0.21
$^5B_{1g}$	$(d_{x^2-y^2})^2(d_{z^2})^1(d_{xz},d_{yz})^2(d_{xy})^1$	0.31	0.20	0.16	0.45	0.46
$^3B_{1g}$	$(d_{x^2-y^2})^1(d_{z^2})^1(d_{xz},d_{yz})^4(d_{xy})^0$	0.41	0.39	0.43	0.48	0.49
$^1A_{1g}(A)$	$(d_{x^2-y^2})^2(d_{z^2})^2(d_{xz},d_{yz})^2(d_{xy})^0$	1.29	1.30	1.30	1.30	1.24
$^1A_{2g}$	$(d_{x^2-y^2})^2(d_{z^2})^2(d_{xz},d_{yz})^2(d_{xy})^0$	1.31	1.31	1.31	1.30	1.29
1E_g	$(d_{x^2-y^2})^2(d_{z^2})^1(d_{xz},d_{yz})^3(d_{xy})^0$	1.34	1.31	1.33	1.34	1.34
$^1A_{1g}(B)$	$(d_{x^2-y^2})^2(d_{z^2})^0(d_{xz},d_{yz})^4(d_{xy})^0$	1.52	1.47	1.52	1.64	1.64

3.4.2. *Results and Discussion.* The results of the CASPT2/RASPT2 calculations performed in this work are presented in Table 6. All reported energies are relative to the $^3A_{2g}$ state, which was found to be the lowest triplet at all calculational levels, with the lowest $^3E_g(A)$ state a 0.08–0.09 eV.

First, a set of CASPT2 calculations was performed based on the CAS(8,11) space described above. As expected, these calculations predict a high-spin, $^5A_{1g}$ ground state (with orbital occupancies shown in Figure 1c). Somewhat surprisingly, however, the $^3A_{2g}$ state is calculated at 0.09 eV, as compared to 0.19 eV in our (latest) previous CASPT2 study.²¹ The difference between these two sets of results is due exclusively to the larger basis sets used in the present calculation.

Next, the CASPT2 calculations based on the CAS(8,11) space were confronted with a set of RASPT2 calculations making use of the same global active space RAS(8,11), however with the five Fe 3d' orbitals placed in RAS3. With the excitation level limited to SD, this active space is noted as (8,0,2;0,6,5). From our previous RASPT2 benchmark study,⁶ we have learned that the 3d double-shell effect may be reasonably well described (i.e., to within 0.1–0.2 eV as compared to the full CASPT2) with the correlating 3d' shell moved from RAS2 to RAS3 and allowing only SD excitations. This finding is confirmed by the present results. Table 6 shows that moving the Fe 3d' shell from RAS2 to RAS3 does not significantly affect the results obtained for the triplet and open-shell singlet states. The closed-shell $^1A_{1g}(B)$ state is calculated 0.05 eV lower with RASPT2 than with CASPT2. However, a larger difference, 0.10 eV, is found for the quintet states. This difference is in line with previous results found for the Ni atom and copper tetrachloride,⁶ where differences of 0.1 eV between CASPT2 and RASPT2 were generally considered acceptable. However, in the present case the consequence is a further increase of the error on the $^5A_{1g}$ – $^3A_{2g}$ splitting at the RASPT2 level.

By putting the 3d' shell in RAS3 we can, however, extend the global size of the RAS space to RAS(34,35), i.e., RAS(8,11) supplemented with the full set of 24 porphyrin π orbitals, containing 26 electrons. The 13 doubly occupied porphyrin π orbitals were put in RAS1, whereas the 11 π^* were added to the Fe 3d' orbitals in RAS3. The results of this set of RASPT2(34,2,2;13,6,16) calculations are compared to the corresponding RASPT2(8,0,2;0,6,5) results in Table 6. Obviously, including the porphyrin π system in the active space does not significantly alter the relative energy of the different Fe 3d' states in this ferrous porphyrin. The three lowest-lying states are hardly affected, whereas for the higher

states, differences of at most 0.05 eV are found between the two sets of results.

As a next option, we decided to run a set of CASPT2 calculations based on an active space of 15 orbitals and including 16 electrons, i.e., the CAS(8,11) space was extended with the Fe semicore (3s,3p) orbitals. It has been known for a long time⁷⁶ that intershell correlation effects between the 3d and the (3s,3p) shells may have a considerable effect on the relative energy of different low-lying states in first-row TM ions and their complexes. However, even if the first study of this effect already indicated that the CASPT2 results for the relative energies of different dⁿ states in TM ions are systematically (slightly) improved when including the semicore shells in the active space, this type of correlation has since then almost invariably been treated at the CASPT2 rather than at the CASSCF level. An obvious reason for this is that making four more orbitals active on top of all other, more important, valence and 3d' orbitals would in many cases result in an active space which is too large to be handled by CASSCF/CASPT2. This is not the case here. Due to the high symmetry of Fe(P), CASPT2(16,15) calculations are perfectly possible. Furthermore, we can make use of this opportunity to test whether the (3s,3p)–3d intershell correlation effects may also be treated accurately by means of RASPT2 rather than CASPT2, i.e., by including the (3s,3p) shells in RAS1 and allowing up-to-doubles into the 3d shell. Relative energies obtained from both sets of calculations, CASPT2(16,15) and RASPT2(16,2,0;4,11,0), are given in Table 6. With exception of one state ($^1A_{1g}(A)$, showing a difference of 0.06 eV), the two sets of results are close to within 0.01 eV, thus confirming that the semicore (3s,3p) orbitals may conveniently be placed in RAS1 rather than in RAS2. Treating the intershell correlation effects between the 3d and the (3s,3p) shells variationally by SD does, however, significantly affect the calculated relative energies of a number of specific states with respect to the $^3A_{2g}$ state, i.e., all quintet states and the closed-shell $^1A_{1g}(B)$ state. As compared to CASPT2(8,11), the energy of the three quintet states is increased by 0.09–0.13 eV in the CASPT2(16,15) treatment, whereas the $^1A_{1g}(B)$ state is shifted upward by 0.12 eV. The triplet and the open-shell singlet states are much less affected. As an important consequence, the $^3A_{2g}$ state is now predicted as the ground state of Fe(P), with both $^5A_{1g}$ and $^3E_g(A)$ low-lying, within 0.1 eV.

The data presented in Table 6 are adiabatic, i.e., they refer to CASPT2/RASPT2 energies obtained for optimized structures (with DFT) for each separate state. These structures are quite different, in particular the Fe–N distances, for the intermediate-spin

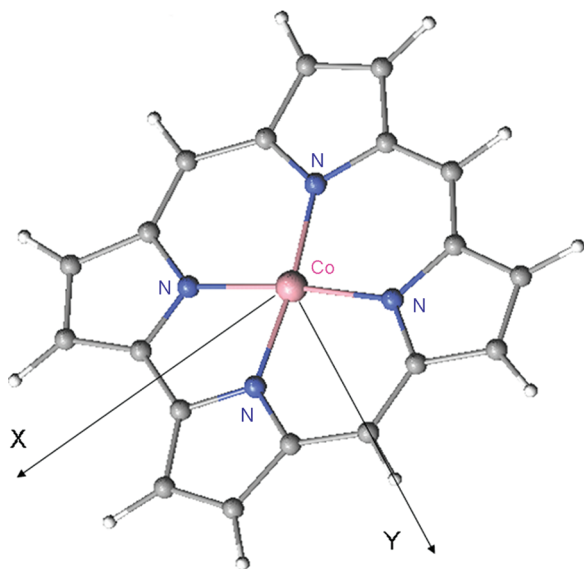


Figure 2. Molecular structure of cobalt corrole.

states ${}^3A_{2g}$, 3E_g ($R(\text{Fe}-\text{N}) = 1.989 \text{ \AA}$ for both states) and the high-spin state ${}^5A_{1g}$ ($R(\text{Fe}-\text{N}) = 2.053 \text{ \AA}$). When calculating instead the vertical electronic spectrum at the ${}^3A_{2g}$ and ${}^5A_{1g}$ structures, respectively, the ${}^5A_{1g}$ is only slightly raised at the ${}^3A_{2g}$ structure, to 0.08 eV, but it becomes the ground state at its own optimal structure, with both triplet states considerably raised in energy, lying at 0.32 eV (${}^3A_{2g}$) and 0.40 eV (3E_g) (with a second quintet state, 5E_g , at 0.17 eV). With three close-lying states at the ${}^3A_{2g}$ structure (but not at the ${}^5A_{1g}$ structure), SOC may become important enough to significantly influence the relative energy between both structures. Moreover, mixing between the three states under SOC may be responsible for the high ground-state magnetic moment, 4.4–4.7 μ_B , observed experimentally for FeTPP and FeOEP.^{55,59} Both premises were explored by treating SOC within the manifold of the seven lowest LF states (all states below 1.0 eV in Table 6), by means of the RAS state interaction (RASSI)-SOC method^{29,77} making use of atomic mean-field integrals (AMFI).^{78–80} The ground-state magnetic moment was calculated making use of the method presented in ref 81. The results of the calculations including SOC are presented in the Supporting Information. At the ${}^3A_{2g}$ structure, the ground state after SOC becomes a mixture of ${}^3A_{2g}$ (68%), 3E_g (13%), and ${}^5A_{1g}$ (18%). The calculated magnetic moment is 4.43 μ_B , in close correspondence with the experimental values of 4.4–4.7 μ_B . This gives further support to the accuracy of the CASPT2(16,15) results for the relative energies of the three low-lying states in Fe(P). At the ${}^5A_{1g}$ structure the SOC wave function is found to be predominantly quintet: ${}^5A_{1g}$ (91%) and 5E_g (5%), with a small (3%) ${}^3A_{2g}$ contribution. A magnetic moment of 5.57 μ_B is calculated for this state, slightly higher than the value for a pure quintet spin, 4.9 μ_B , due to the orbital contribution of the 5E_g state. After SOC, the energy of the ${}^3A_{2g}$ structure is stabilized by 0.06 eV, the ${}^5A_{1g}$ structure by only 0.02 eV. As such, our ‘best’ estimate for the relative energy of both structures becomes 0.08 eV!

3.5. Low-Lying States of Cobalt Corrole. Corrole is a porphyrinoid compound in which one of the meso carbon atoms has been removed and replaced by a direct pyrrole–pyrrole bond. The remaining three meso positions remain occupied by carbon atoms. When fully deprotonated, corrolate becomes a

trianionic ligand. The smaller trianionic corrolate ligand has a greater ability to stabilize higher central metal oxidation states than the larger dianionic porphyrinato ligand. On the other hand, the corrolate ligand is also easily oxidizable and may transfer one of its electrons to the central metal, e.g., in nickel and copper corrole,^{9,82,83} to become itself ‘noninnocent’, i.e., a dianion radical. These special properties of corrolate have driven an intense investigation of TM corrole complexes.

Many experimental and theoretical studies have contributed to the characterization of cobalt corrole systems.^{84–92} Based on the magnetic moment (3.2 μ_B) and ${}^1\text{H}$ NMR spectra (showing strong paramagnetic shifts) the Co(OEC) (OEC = octaethylcorrole) system was characterized as an intermediate-spin Co(III)($S = 1$) center coordinated by a corrolate³⁻ ligand in a square planar coordination environment.⁸⁴ The corrole was in this system clearly indicated as an innocent ligand. As such, the cobalt(d^6) atom in Co-corrole is isoelectronic with the iron(d^6) atom in Fe-porphyrin, and one would a priori expect a similar electronic structure. An early spin-unrestricted DFT (BP86 functional) study of the electronic structure of unsubstituted Co-corrole confirmed that the electronic ground state indeed also has intermediate-spin state ($S = 1$), corresponding to an electronic configuration $(d_{x^2-y^2})^2(d_{z^2})^2(d_{xz}, d_{yz})^2$.⁹³ A low-lying singlet state is found at 0.26 eV, whereas the lowest quintet state is calculated at 1.08 eV. All three states were found to have the same (+3) oxidation state on cobalt. The lowest singlet state is an open-shell state corresponding to the same configuration as the triplet ground state (cf. the lowest singlet state in iron porphyrin; see previous section). However, unlike iron porphyrin, the lowest quintet state is in Co-corrole much higher-lying and does not correspond to a Co(III) high-spin configuration. Instead, a $\pi \rightarrow \pi^*$ transition is found to give rise to the lowest quintet excited state. That the σ antibonding $3d_{xy}$ orbital does not easily become occupied in Co-corrole may be explained by the rigid trianionic-ligand framework of corrolate, with a smaller cavity and corresponding stronger σ donation than the porphyrinate ligand in Fe-porphyrin.

In this work we report the results obtained from a comparative CASPT2 versus RASPT2 study of the ground state and a selected set of low-lying excited states of unsubstituted cobalt corrole, denoted as Co(C) (see Figure 2 for the structure (C_{2v}) and orientation of the axes). For the ground state, both CASPT2 and RASPT2 confirm the previous DFT prediction of a triplet Co(III) state 3B_1 with an electronic configuration of $(d_{x^2-y^2})^2(d_{z^2})^2(d_{xz})^1(d_{yz})^1$. Rather than considering other Co(III) states we have focused here on the possible occurrence of low-lying Co(II)-(C)²⁻ radical states. Starting from the 3B_1 ground state, we have considered excitations out of the two highest corrolate π orbitals of symmetry a_2 and b_2 into the cobalt $3d_{xz}$ orbital, giving rise to either ferro- or antiferromagnetic coupling between the remaining singly occupied cobalt $3d_{yz}$ orbital (residing in the representation a_2) and an a_2 or b_2 corrole π -radical. Of course, other low-lying Co(II) states, corresponding to different $(d_{x^2-y^2})^2(d_{xz}, d_{yz}, d_{z^2})^5(C)\pi^1$ arrangements, are also possible. A more complete description of the Co(C) spectrum will be provided in a separate paper.

3.5.1. Computational Details. Single point CASPT2/RASPT2 calculations were performed on structures obtained from DFT (in its unrestricted formalism, employing the PBE0 functional) using TZVP basis sets on all atoms. C_{2v} symmetry was maintained in all geometry optimizations. For conformity with iron porphyrin (previous section) the molecule was placed in the xy -plane with the x -axis as the C_2 axis (Figure 1).

Table 7. Relative Energy (eV) of Selected Low-Lying Triplet Electronic States of Co(C) with Respect to the 3B_1 Ground State.^a

CI based on CASSCF orbitals					
	CASPT2	SD	SDT	SDTQ	full CI
	RASPT2(12, <i>l</i> , <i>m</i> ;4,4,6)				
$^3B_1((d_{yz})^1(b_2)^1)$	0.29	0.43	0.37	0.36	0.36
$^3A_1((d_{yz})^1(a_2)^1)$	0.78	0.95	0.88	0.87	0.86
	RASPT2(12, <i>l</i> , <i>m</i> ;2,6,6)				
$^3B_1((d_{yz})^1(b_2)^1)$	0.29	0.36	0.29	0.28	0.28
$^3A_1((d_{yz})^1(a_2)^1)$	0.78	0.88	0.80	0.78	0.78
	RASPT2(12,0, <i>m</i> ;0,8,6)				
$^3B_1((d_{yz})^1(b_2)^1)$	0.29	0.35	0.29	0.28	0.28
$^3A_1((d_{yz})^1(a_2)^1)$	0.78	0.87	0.79	0.78	0.78
CI based on RASSCF/SD orbitals					
	CASPT2	SD	SDT	SDTQ	full CI
	RASPT2(12, <i>l</i> , <i>m</i> ;4,4,6)				
$^3B_1((d_{yz})^1(b_2)^1)$	0.29	0.42	0.37	0.36	0.36
$^3A_1((d_{yz})^1(a_2)^1)$	0.82	0.98	0.92	0.90	0.90
	RASPT2(12, <i>l</i> , <i>m</i> ;2,6,6)				
$^3B_1((d_{yz})^1(b_2)^1)$	0.33	0.38	0.32	0.31	0.31
$^3A_1((d_{yz})^1(a_2)^1)$	0.85	0.92	0.85	0.84	0.84
	RASPT2(12,0, <i>m</i> ;0,8,6)				
$^3B_1((d_{yz})^1(b_2)^1)$	0.33	0.37	0.32	0.31	0.31
$^3A_1((d_{yz})^1(a_2)^1)$	0.85	0.91	0.85	0.84	0.84

^a Single point CASSCF/CASPT2 and RASSCF/RASPT2 (CI only) calculations.

The active space used in the CASSCF/CASPT2 calculations was constructed similarly as for the iron porphyrin presented in the previous section, i.e., with eight electrons distributed over the cobalt 3d, 3d', and the bonding $\sigma(\text{Co-N})$ orbital. However, since the 3d_{xy} orbital is in this case unoccupied in all considered states, its correlating 3d'_{xy} orbital was not included. This gives a CAS(8,10) space. On top of this, to consider also Co(II)-(C)^{•2-} radical states the two highest-occupied π orbitals of the corrole (a_2 , b_2) and their correlating π^* orbitals were made active, leading to a CAS(12,14) space.

RASPT2 calculations were performed with three global active spaces: the original (12,14) space and two larger spaces, constructed by extending this (12,14) space either with all or with a selection of the remaining corrole π and π^* orbitals, giving respectively a RAS(34,33) or RAS(26,25) space. The latter active space lacks the four occupied corrole π and their four correlating π^* orbitals built as combinations of the 2p _{π} orbitals of the eight β carbons of the corrole rings (a plot of these orbitals can be found in the Supporting Information). A similar active space was employed in a recent RASPT2 study of the iron-oxo porphyrins FeO(P)⁺ and FeO(P)Cl.¹⁰ The presence of an axial oxo ligand in these complexes put higher demands on the size of the RAS2 space (which has to contain the six orbitals involved in the very covalent FeO bond), thus precluding the use of a RAS space including all porphyrin π and π^* orbitals. The present test calculations with the smaller RAS(26,25) therefore serve as a

(additional) check for the quality obtained from the RASPT2 calculations on these iron-oxo systems.

For the Co(C) calculations in this work, three different subpartitions of the global RAS space were considered. In all three cases, the four Co 3d' orbitals and all active π^* orbitals were collected in RAS3, which thus consists of 6 orbitals in all RAS(12,14) calculations, 10 orbitals in all RAS(26,25) calculations, and 14 orbitals in all RAS(34,33) calculations. Also, the two singly occupied orbitals in each of the open-shell Co(II) π -radical cation states were systematically put in RAS2, thus avoiding the strict need of including up-to-triple excitations out of RAS1 and into RAS3, which would become computationally too demanding for the two largest active spaces. Next to these two state-specific orbitals, the smallest RAS2 space only contains the bonding and antibonding orbitals involved in the covalent Co 3d_{xy}-corrole σ interaction. The second option for RAS2 is to include all four orbitals that become singly occupied in any of the five considered states, i.e., corrole π HOMO (a_2 , b_2), and the cobalt 3d_{xz}, 3d_{yz} orbitals. Adding also the (σ, σ^*) couple gives a six-orbital RAS2. The largest RAS2 considered includes on top of the two remaining, doubly occupied, Co 3d_{z²}, 3d_{x²-y²} orbitals.

3.5.2. Results and Discussion. A first series of RASPT2 calculations was performed based on the (12,14) active space, and considering only triplet states: the Co(III) ground state $^3B_1((d_{xz})^1(d_{yz})^1)$ and the two excited states 3B_1 , 3A_1 with configuration $((d_{xz})^2(d_{yz})^1)$ on Co(II) coupled ferromagnetically to either a corrole b_2 or a_2 π radical. The goal of these smaller calculations is two-fold: (a) to explore the minimum size of the RASSCF reference wave function that would be needed to obtain an accuracy which is close to the parent CASPT2 result, by systematically varying the size of RAS2 and the excitation level l , m , and (b) to investigate the effect of the optimization level of the active orbitals on the relative energies obtained from CASPT2/RASPT2. A similar exercise was already performed for CrF₆ in Section 3.1. Here, we decided to investigate this point further by comparing the results from two sets of CASPT2 and RASPT2 calculations, starting from reference wave functions in which only the CI coefficients are optimized (allowing up to full CI) but making use of orbitals obtained from either the full CAS(12,14) calculation or from the corresponding RAS(12,2,2;4,4,6), RAS(12,2,2;2,6,6), or RAS(12,0,2;0,8,6) calculation. The results of these RASPT2(12,14) calculations are shown in Table 7. As can be seen from this table, the differences between the calculated relative energies with the two sets of orbitals are relatively small, 0.06 eV or less. This further confirms that RASSCF orbitals obtained from an SD calculation may rather safely be used to obtain reference wave functions with a higher excitation level, without a strict need for further (time-consuming and generally poorly convergent) orbital optimizations.

Turning next to the different excitation levels in Table 7, we first note that the full CI reference wave function of the RASPT2 results shown in the rightmost column of this table is equal to the CASSCF wave function giving rise to the CASPT2 data in the leftmost column (i.e., both wave functions are built from the same orbitals and contain exactly the same number of configuration state functions). Any difference between both sets of results therefore should be traced back to the different perturbational treatment and more specifically to the neglect in RASPT2 of the off-diagonal Fock matrix elements connecting active orbitals belonging to different RAS subspaces. For a more detailed discussion of this phenomenon, see Section 3.2.

Table 8. Relative Energy (eV) of Selected Low-Lying Electronic States of Co(C) with Respect to the 3B_1 Ground State

	$^3B_1((d_{yz})^1(b_2)^1)$	$^1B_1((d_{yz})^1(b_2)^1)$	$^1A_1((d_{yz})^1(a_2)^1)$	$^3A_1((d_{yz})^1(a_2)^1)$
CASPT2(12,14)	0.29	0.29	0.43	0.78
RASPT2(12,2,2;4,4,6)	0.42	0.39	0.67	0.98
RASPT2(12,3,3;4,4,6)	0.37	0.34	0.60	0.92
RASPT2(12,2,2;2,6,6)	0.38	0.35	0.61	0.92
RASPT2(12,3,3;2,6,6)	0.32	0.29	0.53	0.85
RASPT2(26,2,2;11,4,10)	0.61	0.63	0.70	0.90
RASPT2(26,3,3;11,4,10)	0.60	0.62	0.69	0.89
RASPT2(26,2,2;9,6,10)	0.61	0.63	0.65	0.86
RASPT2(26,3,3;9,6,10)	0.58	0.60	0.59	0.81
RASPT2(34,2,2;15,4,14)	0.60	0.62	0.63	0.80
RASPT2(34,2,2;13,6,14)	0.67	0.70	0.69	0.86

The differences between the RASPT2 allowing up to full CI and CASPT2 results are significant, 0.07–0.08 eV, only for the calculations with the smallest RAS2, where the two sets of orbitals ($3d_{xz}$, $3d_{yz}$) and the corrole π HOMO orbitals of the same symmetry (b_2 , a_2) are divided over RAS1 and RAS2. After including both sets of orbitals in RAS2, the RASPT2 calculations almost exactly converge (with increasing excitation level) to the CASPT2 results. The differences between the RASPT2 results obtained with the six- and eight-orbital RAS2 (the latter containing two extra cobalt 3d orbitals) are insignificantly small at all excitation levels. With these two RAS2 spaces, the difference between the CASSCF (or full CI) and the more limited RASSCF wave functions essentially comes down to a less complete description of the 3d double-shell effect and of corrole $\pi-\pi^*$ correlation in the latter wave function. As can be seen, up-to-triple RAS2 \rightarrow RAS3 excitation is necessary but sufficient to bring the RASPT2 results to within 0.02 eV of the corresponding full CI RASPT2 results. The difference between the SD and full CI results is of the order of 0.1 eV. Differences of the same order of magnitude were also found for the energy differences between different (d^6) spin states in Fe(P) (Section 3.4), and will also be met in the next section, when trying to describe the relative energy of different low-lying states in FeO and FeO $^-$.

From the results in Table 7, it may therefore be concluded that it takes a six-orbital RAS2 combined with SDT excitations to obtain RASPT2 results that essentially match CASPT2. Unfortunately, the combination of these two conditions gives rise to very large RASSCF wave functions when extending the size of the global RAS space to more than twenty or so orbitals.

In Table 8 we compare the RASPT2 results obtained with the three global RAS spaces for both the ferromagnetic 3B_1 and 3A_1 and antiferromagnetic 1B_1 and 1A_1 Co(II)(C) $^{2-}$ states. Both the four- and six-orbital RAS2 were used, allowing up-to-triple excitations for the two smallest RAS spaces but only up-to-doubles for the largest global RAS space. Comparing first the data with the two largest RAS spaces, we find differences of at most 0.1 eV between the numbers obtained with the same RAS2. The effect of triples is, for the global RAS(26,25) space, of the same order of magnitude, around 0.05 eV. On the whole, the combination of all data collected in both tables for Co(C) indicate that, for complicated electronic situations, such as the present cobalt corrole system with its different metal oxidation states in an extended π -delocalized ligand system, computational limitations still prevent the full exploitation of the possibilities of RASPT2, and the compromises that have to be made in the construction of the RAS space may lead to errors of around 0.2 eV.

This, however, does not alter the fact that very valuable results may be obtained with this method, pointing to deficiencies in the CASPT2 method that are the result of too limited active spaces. Two such deficiencies are eminent from the results in Table 8. First, from a comparison between the results obtained with the (12,14) and the more extended active spaces, it is clear that the relative energy of the a_2 versus b_2 π -radical states is not well described when including in the active space only the ‘Gouterman’ set of two π and two π^* orbitals. The energy of the $^3,^1A_1$ (a_2 radical) states remains virtually constant when extending the active space with extra π orbitals. However, the $^3,^1B_1$ states (b_2 radical) are quite strongly (about 0.25 eV) shifted upward in the RASPT2 calculations with more extended active spaces. We believe that this failure of CASPT2 to correctly reproduce the splitting between the a_2 and b_2 radical states should be brought back to an unbalanced treatment of corrole $\pi-\pi^*$ correlation when only four π orbitals are active. In the Co(III) ground state, all π orbitals are doubly occupied, and those two (π, π^*) pairs that give rise to the largest correlation energy enter the active space. However, these are not (necessarily) also the HOMO π orbitals, i.e., the ones that are most easily depopulated and therefore become active in the open-shell π -radical states. A plot of the four active corrole π orbitals in the CAS(12,14) treatment of the three triplet states is presented in Figure 3. One can see that these orbitals indeed quite strongly change shape between the different states considered.

A second failure of CASPT2 concerns the description of the magnetic coupling in the Co(II) π -radical states. Two cases should be distinguished here. In the $^3,^1B_1$ states, the coupling occurs between two electrons residing in orbitals that are mutually orthogonal: $b_2(d_{xz})$ and $a_2(\pi)$. As the orbitals are strictly orthogonal, the coupling should be weakly ferromagnetic. As can be seen from Table 8, this is corroborated by the RASPT2(26,25) and RASPT2(34,33) results, with the 3B_1 state lying below 1B_1 by 0.02–0.03 eV. On the other hand, making use of the smaller (12,14) space, both CASPT2 and RASPT2 incorrectly predict the coupling to be weakly antiferromagnetic. In the second case, the a_2 π -radical states $^3,^1A_1$, the interaction occurs between two orbitals that are weakly overlapping. As such, antiferromagnetic coupling is expected and confirmed by all results in Table 8. However, here the absolute value of the magnetic coupling is overestimated quite strongly with CASPT2, predicting a $^3A_1-^1A_1$ energy gap of more than 0.3 eV. This gap is decreased to 0.2 eV or less when extending the active space with extra corrole π^* orbitals.

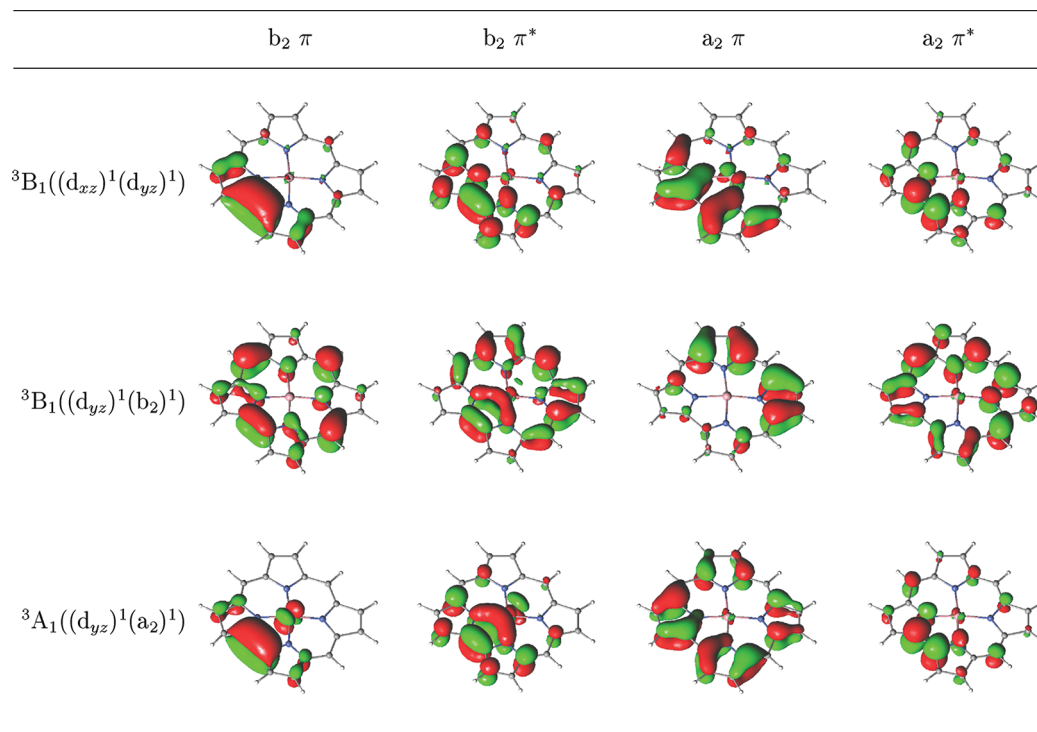


Figure 3. Four HOMO–LUMO π orbitals (as obtained from the CAS(12,14) calculations) for the selected triplet Co(Cor) states. The contour values are $\pm 0.04 \text{ e/au}^3$.

As a final note, we should also mention that the same two obvious failures of CASPT2 were already observed before when describing the relative energy of different electromers of copper corroles⁹ and of iron–oxo porphyrins.¹⁰ Obviously, even with its non-negligible error bars, the RASPT2 method has clear advantages over standard CASPT2 calculations also for these complicated cases.

3.6. Double-Shell Effects in the Low-Lying States of FeO and FeO[−]. The iron–oxygen bond constitutes an important aspect of biological transportation and catalysis. This has been the prime motivation for many experimental and theoretical studies about the electronic structures of small iron oxides clusters in the gas phase.^{23,94–100} The electronic structure of iron monoxide and its singly charged anion is particularly challenging because of the high density of low-lying states in both systems.^{23,100} A theoretical multiconfigurational approach based on the CASPT2 method was used only recently to describe the electronic structure of FeO/FeO[−] and to provide an interpretation of the photoelectron spectrum of the anion starting from a ${}^6\Sigma^+$ ground state.²³ However, this CASPT2 assignment was shown to contradict the high-resolution autodetachment spectrum of FeO[−] which indicates that rather the ${}^4\Delta$ state is the ground state of the anion.¹⁰⁰ In a very recent ab initio study, high-level single-reference (RCCSD(T)) and multireference (MRCI) results were presented, confirming the CASPT2 prediction that the ${}^6\Sigma^+$ state is situated below the ${}^4\Delta$ state.¹⁰¹ However, as the calculated energy difference between both states is extremely small (0.05 eV with R-CCSD(T), including Fe (3s,3p) correlation and scalar relativistic effects), it is proposed that ${}^6\Sigma^+$ could be embedded in the SOC components of ${}^4\Delta$.

In a qualitative orbital scheme for FeO the oxygen 2p orbitals give rise to bonding doubly occupied molecular orbitals, while the remaining six valence electrons are distributed among the

predominantly iron 3d and 4s orbitals. Of these, the two lowest lying levels, 9σ and 1δ , are nonbonding, while the higher-lying 4π orbitals and in particular the 10σ orbital are antibonding. Occupying all six valence orbitals by one electron with parallel spin results in a ${}^7\Sigma^+$ ($9\sigma^1 1\delta^2 4\pi^2 10\sigma^1$) state. By transferring an electron from the most strongly antibonding 10σ orbital to the nonbonding 9σ or 1δ levels the ${}^5\Sigma^+$ ($9\sigma^2 1\delta^2 4\pi^2 10\sigma^0$) and ${}^5\Delta$ ($9\sigma^1 1\delta^3 4\pi^2 10\sigma^0$) states are reached. The latter state is generally considered as the ground state of FeO. For the corresponding FeO[−] anion, this orbital picture leads to three possible low-lying electronic states obtained by placing the extra electron in one of the nonbonding orbitals 9σ or 1δ . Starting from the neutral ${}^7\Sigma^+$ state, this gives ${}^6\Sigma^+$ ($9\sigma^2 1\delta^2 4\pi^2 10\sigma^1$) or ${}^6\Delta$ ($9\sigma^1 1\delta^3 4\pi^2 10\sigma^1$) states, while both quintets may accept the extra electron to form the same low-lying ${}^4\Delta$ ($9\sigma^2 1\delta^3 4\pi^2 10\sigma^0$) state.

The original CASPT2 calculations²³ were performed with an active space of 14 orbitals, consisting of nine valence orbitals O 2p, Fe (3d,4s) and extended with five virtual orbitals of σ , π , δ symmetry. Although initially intended to describe the Fe 3d double-shell effect, the extra σ , π orbitals turned out to have strongly mixed Fe 3d/O 3p character, the latter even providing the most dominant contribution. A more balanced description should include both the double-shell effect on iron and the oxygen anion. This would then give rise to an active space of 17 orbitals, accommodating either 12 (FeO) or 13 (FeO[−]) electrons. CASPT2 calculations with such a large active space are computationally expensive, in particular if they have to be combined with the optimization of geometrical parameters. Therefore we decided to investigate instead the possibilities of RASPT2, using an active space in which the double-shell effect on both atoms is described by an active space with the appropriate correlating orbitals placed in the RAS3 subspace. Moreover, given the option for a larger active space in RASSCF, we also chose to include

Table 9. Fe–O Bond Distance (Å) and Relative Energy (eV) with Respect to the ${}^6\Sigma^+$ state of FeO^- of the Different Electronic States of FeO and FeO^-

	FeO^-			FeO		
	${}^4\Delta$	${}^6\Sigma^+$	${}^6\Delta$	${}^5\Delta$	${}^5\Sigma^+$	${}^7\Sigma^+$
Fe–O Bond Distance (Å)						
CASPT2(12/13,14) ^a	1.634	1.683	1.688	1.612	1.626	1.677
RASPT2(14/15,0,2;0,10,19)	1.619	1.677	1.673	1.611	1.620	1.671
Relative Energy (eV)						
CASPT2(12/13,14) ^a	0.13	0.00	0.36	1.51	1.55	1.75
RASPT2(14/15,0,2;0,10,9)	0.04	0.00	0.29	1.24	1.32	1.64
RASPT2(14/15,0,3;0,10,9)	0.01	0.00	0.34	1.33	1.41	1.73
RASPT2(14/15,0,4;0,10,9)	0.02	0.00	0.23	1.39	1.48	1.77
RASPT2(14/15,0,5;0,10,9)	0.02	0.00	0.23	1.40	1.49	1.78
RASPT2(14/15,0,6;0,10,9)	0.02	0.00	0.23	1.40	1.49	1.78

^a From Hendrickx and Anam,²³ CASPT2 results obtained with ANO-rcc basis sets contracted to $\text{Fe}[8s7p6d4f2g1h] \text{O}[7s6p4d3f1g]$, except for the ${}^6\Delta$ state, for which only results with a smaller contraction $\text{Fe}-[8s7p5d4f2g]$ and $\text{O}[6s5p4d2f]$ were reported.

the O (2s,3s) shells, thus obtaining a RAS(14/15,0,m;0,10,9) including all FeO valence orbitals in RAS2 and their correlating shells in RAS3. Calculations with an excitation level $m = 2-6$ were performed.

The molecular orbitals and the Fe–O bond distance were optimized only for the SD calculations, while single-point CI calculations were performed for the higher excitation levels, employing the orbitals and Fe–O distances from the optimized SD calculations. Calculations were performed with ANO-rcc basis sets,^{25,26} using a very extended contraction scheme: $[10s9p8d6f4g2h]$ for iron and $[8s7p4d3f2g]$ on O. The results obtained from this work are presented in Table 9, where they are compared to the results from our previous CASPT2 study.²³

Considering first the calculated Fe–O bond distances, one can observe a significant bond shortening when going from CASPT2 to RASPT2. The largest differences, 0.015 Å, are found for the FeO^- ${}^4\Delta$, ${}^6\Delta$ states. The Fe–O distances obtained from RASPT2 show a very clear pattern: Very similar distances, 1.671–1.677 Å, are obtained for the high-spin states of both molecules, whereas the calculated distances for all low-spin states are considerably shorter, 1.611–1.620 Å, but also these values differ by less than 0.01 Å among themselves. The reason for this observation lies in the occupation of the strongly antibonding 10σ orbital in the high-spin states, and the fact that among themselves, both the low- and high-spin states only differ in the occupation of the 9σ and 1δ nonbonding orbitals. A similar pattern is also found at the CASPT2 level, although the differences are here somewhat larger.

Table 9 also shows that the different low-lying states of FeO and FeO^- are indeed close in energy, lying within a range of 0.4 eV for each of the methods used. For FeO, all calculations confirm the generally accepted assignment of a ${}^5\Delta$ ground state. At the CASPT2 level the ${}^6\Sigma^+$ state was calculated as the ground state of FeO^- . This assignment was, however, disputed¹⁰⁰ based on its experimental autodetachment spectrum,⁹⁴ showing a fully rotationally resolved fine structure that can only originate from a ground state with the spin–orbit fine structure characteristic of a ${}^4\Delta$ state. As one can see from Table 9, the energy difference

between the two candidate ground states of FeO^- is substantially reduced when introducing a more complete description of the double-shell effect in the RASPT2 calculation. A similar trend is found for all other low-spin states, which are all stabilized by around 0.1 eV in RASPT2 as compared to CASPT2. On the other hand, the relative energy between the two high-spin states, ${}^7\Sigma^+ - {}^6\Sigma^+$ is slightly higher with RASPT2(SDTQ) than with CASPT2. Convergence of the RASPT2 results is essentially reached at the SDTQ level. It should be noted though that the differences between the lower excitation levels in RASPT2 are quite significant, up to 0.16 eV between SD and SDTQ, and show oscillating behavior with the excitation level in case of the ${}^6\Delta$ state. Similar (although generally smaller) errors in the RASPT2-(SD) description of the Fe 3d double-shell effect were already met in case of Fe(P) (Table 6). It should be clear from these and other RASPT2 results in this work (e.g., the test results for Co(C) in Tables 7 and 8) that uncertainties of 0.1–0.2 eV on relative energies are to be expected from RASPT2 when limiting the excitation levels out of RAS1 and into RAS3 to only SD. Such error bars become particularly bothersome when trying to describe the relative energy of close-lying states, as is the case here.

An important observation from the results in Table 9 is also that the two states ${}^4\Delta$ and ${}^6\Sigma^+$ in FeO^- are very nearly degenerate. It is quite probable that, when including SOC in the calculations, the actual ground state of FeO^- might in fact become a mixture of both states, with an Fe–O distance which is intermediate between the calculated distances of 1.619 Å for the ${}^4\Delta$ and 1.677 Å for the ${}^6\Sigma^+$ states. Such a mixed ground state is in fact suggested by the experimental Fe–O bond length of 1.641 Å (derived from the experimental rotational constant for the $\Omega = 7/2$ ground state observed in the autodetachment spectrum) and will be further investigated in a separate study, by performing relativistic RASPT2 geometry optimizations, including SOC for all low-lying states of FeO^- .

4. SUMMARY AND CONCLUSIONS

In this work we have presented the results of an explorative study of different properties of a series of six representative first-row TM systems by means of the recently introduced RASPT2 method.⁴ Different RASPT2 strategies were investigated for each molecule, aimed at obtaining the highest possible accuracy to be reached by a multiconfigurational perturbation theory method at the lowest possible computational cost. This was accomplished by a two-step procedure, starting first with a series of calculations with a ‘standard’ active space, containing the most important correlation effects (3d double-shell and/or correlation effects connected to covalent TM–ligand interactions) but with a size which is small enough to be tractable by CASPT2. In this series of calculations, different RAS subpartitions and excitation levels were explored, aiming at keeping the accuracy of the RASPT2 results as close as possible to CASPT2, while reducing the number of configuration-state functions in the reference wave function. From these calculations, the following general rules for the construction of RAS spaces may be formulated:

- In case of covalent metal–ligand interactions all bonding–antibonding orbital pairs between the ligands and the metal 3d orbitals should be kept together in RAS2, in order to allow full rotational freedom between these orbital pairs in the diagonalization of the active part of the Fock matrix.

- (b) The 3d double-shell effect may be adequately described by putting the second 3d' shell in RAS3. However, in order to avoid errors of the order of 0.1 eV (a few kcal/mol) at least up-to-triple excitations into RAS3 might be necessary in many cases, such as the description of spin state energetics (in Fe(P)), charge-transfer states (in Co(C)), etc. (see also the cases of Ni and CuCl_4^{2-} in a previous RASPT2 study⁶). In cases where combined double-shell effects are described by means of RASSCF, such as the case of FeO in the present work (see also the binuclear copper compounds described previously)⁴ up-to-quadruple excitations into RAS3 are necessary.
- (c) Excitation processes may be described either by using a different RAS2 for each specific state, containing those orbital(s) that become singly occupied [on top of the covalent orbitals described in (a)] or by keeping such orbitals in either RAS1 or RAS3 but then allowing at least up-to-triples in or out of these subspaces, so as to describe the combined effect of excitation/correlation.

In a second step, extensions of the active space (either in RAS1 and RAS3 or both) were considered, aiming at increasing the accuracy of the RASPT2 results beyond the limits of CASPT2. For ferrocene and $\text{Cr}(\text{CO})_6$, the additional orbitals served to provide a more complete description of the double-shell effect as well as correlation within the ligand π orbitals. Including these orbitals in the active space was shown to give a significant improvement of the results obtained from the perturbational treatment, bringing the RASPT2 result for the heterolytic binding energy of ferrocene (641.4 kcal/mol) within the error limits of the experimental value (641 ± 6 kcal/mol), and the RASPT2 excitation energies of the two allowed CT transitions in $\text{Cr}(\text{CO})_6$ to within 0.1 eV of the experimental band maxima. For both molecules, this was accomplished by RASPT2(SD) after constructing the RAS space according to the rules given above. For CrF_6 it turned out that even with a 10-orbital RAS2 at least RASPT2(SDTQ) is necessary to obtain accurate results for the energy difference between the D_{3h} and O_h structures. This may not come as a surprise, given the very strong correlation effects in this molecule involving also excitations from nonbonding F 2p orbitals (in RAS1) into Cr 3d (in RAS2).¹⁴ Nevertheless, also for CrF_6 the RASPT2 method offers a clear improvement over CASPT2, which is dramatically suffering from limitations in the size of the CAS space. For the two heme or heme-related systems Fe(P) and Co(C), the largest active spaces comprised the entire set of heme π orbitals. Including at least 15 corrole π orbitals rather than just a few frontier orbitals (such as the Gouterman set of two π and two π^* orbitals) is shown to give a considerable improvement of the RASPT2 results for the splitting between the a_2 and b_2 radicals on the corrole ligand and for the strength of the magnetic coupling in both types of Co(II) π -radical states. However, with an active space containing as much as 33 orbitals, combining an optimal RAS2 with RASPT2(SDT) becomes computationally unfeasible. Although the RASPT2 results are also for Co(C) superior to CASPT2, a larger error bar of 0.1–0.2 eV has to be accepted for such a difficult case. On the other hand, for Fe(P) including the (full set of) porphyrin π orbitals in the active space does not affect the spin state energetics of the different Fe(II) states to any significant extent. Here it is shown that including instead in the active space the Fe semicore (3s,3p) orbitals significantly improves the description of (3s,3p)–3d intershell correlation effects, thereby correctly

reproducing the $^3A_{2g}$ state as the ground state of Fe(P). The (experimentally observed) high magnetic moment associated with this triplet state is correctly reproduced after introducing SOC in the CASPT2(16,15) calculations, thus allowing mixing of $^3A_{2g}$ with the close-lying $^5A_{1g}$ and 3E_g states. Finally, for FeO^- the present RASPT2 results predict a quasidegenerate $^6\Sigma^+ - ^4\Delta$ ground state (the latter state situated at only 0.02 eV above the former). Also here, SOC may be expected to become crucial in correctly describing the structure and character of the ground state. A more detailed analysis of the low-lying states of FeO/FeO^- will be presented in a separate paper.

In conclusion, the results presented in this work have clearly illustrated the strength of the RASPT2 method for the prediction of several important properties in first-row TM systems, thus establishing this method as a valuable tool for studying large TM complexes with complicated electronic structures, creating the need for very large active spaces.

■ ASSOCIATED CONTENT

S Supporting Information. Contour plots of the active orbitals included in the most extended active space of all molecules. Contour plots of the eight active orbitals that are included in the (34,33) active space of Co(C) but not in the (26,25) space. Results obtained from the calculations on Fe(C) including SOC. This material is available free of charge via the Internet at <http://pubs.acs.org>.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: kristin.pierloot@chem.kuleuven.be.

■ ACKNOWLEDGMENT

This investigation has been supported by grants from the Flemish Science Foundation (FWO) and from the Concerted Research Action of the Flemish Government (GOA).

■ REFERENCES

- (1) Andersson, K.; Malmqvist, P.-Å.; Roos, B. O.; Sadlej, A. J.; Wolinski, K. *J. Phys. Chem.* **1990**, *94*, 5483–5488.
- (2) Andersson, K.; Malmqvist, P.-Å.; Roos, B. O. *J. Chem. Phys.* **1992**, *96*, 1218–1226.
- (3) Ghosh, A. *J. Biol. Inorg. Chem.* **2011**, *16*, 819–820.
- (4) Malmqvist, P.-Å.; Pierloot, K.; Shahi, A. R. M.; Cramer, C. J.; Gagliardi, L. *J. Chem. Phys.* **2008**, *128*, 204109.
- (5) Huber, S. M.; M.Shahi, A. R.; Aquilante, F.; Cramer, C.; L. Gagliardi *J. Chem. Theory Comput.* **2009**, *5*, 2967–2976.
- (6) Sauri, V.; Serrano-Andrés, L.; Shahi, A. R. M.; Gagliardi, L.; Vancoillie, S.; Pierloot, K. *J. Chem. Theory Comput.* **2011**, *7*, 153–168.
- (7) Ruipérez, F.; Aquilante, F.; Ugalde, J. M.; Infante, I. *J. Chem. Theory Comput.* **2011**, *7*, 1640–1646.
- (8) Shahi, A. R. M.; Cramer, C. J.; Gagliardi, L. *Phys. Chem. Chem. Phys.* **2009**, *11*, 10964–10972.
- (9) Pierloot, K.; Zhao, H.; Vancoillie, S. *Inorg. Chem.* **2010**, *49*, 10316–10329.
- (10) Radoń, M.; Broclawik, E.; Pierloot, K. *J. Chem. Theory Comput.* **2011**, *7*, 898–908.
- (11) Andersson, K.; Roos, B. O. *Chem. Phys. Lett.* **1992**, *191*, 507–514.
- (12) Roos, B. O.; Andersson, K.; Fülcher, M. P.; Malmqvist, P.-Å.; Serrano-Andrés, L.; Pierloot, K.; Merchán, M. In *Advances in Chemical Physics*:

New Methods in Computational Quantum Mechanics; Prigogine, I., Rice, S. A., Eds.; John Wiley & Sons: New York, 1996; Vol. XCIII, pp 219–332.

(13) Pierloot, K. Nondynamic Correlation Effects in Transition Metal Coordination Compounds. In *Computational Organometallic Chemistry*; Cundari, T. R., Ed.; Marcel Dekker, Inc.: New York, 2001; pp 123–158.

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

(15) Pierloot, K. *Int. J. Quantum Chem.* **2011**, *111*, 3291–3301.

(16) Veryazov, V.; Malmqvist, P.; Roos, B. O. *Int. J. Quantum Chem.* **2011**, *111*, 3329–3338.

(17) Pierloot, K.; Roos, B. O. *Inorg. Chem.* **1992**, *31*, 5353–5354.

(18) Persson, B. J.; Roos, B. O.; Pierloot, K. *J. Chem. Phys.* **1994**, *101*, 6810–6821.

(19) Pierloot, K.; Persson, B. J.; Roos, B. O. *J. Phys. Chem.* **1995**, *99*, 3465–3472.

(20) Pierloot, K.; Tsokos, E.; Vanquickenborne, L. G. *J. Phys. Chem.* **1996**, *100*, 16545–16550.

(21) Radoń, M.; Pierloot, K. *J. Phys. Chem. A* **2008**, *112*, 11824–11832.

(22) Vancoillie, S.; Zhao, H.; Radoń, M.; Pierloot, K. *J. Chem. Theory Comput.* **2010**, *6*, 576–582.

(23) Hendrickx, M. F. A.; Anam, K. R. *J. Phys. Chem. A* **2009**, *113*, 8746–8753.

(24) Aquilante, F.; De Vico, L.; Ferré, N.; Ghigo, G.; Malmqvist, P.-Å.; Neogrády, P.; Pedersen, T. B.; Pitoňák, M.; Reiher, M.; Roos, B. O.; Serrano-Andrés, L.; Urban, M.; Veryazov, V.; Lindh, R. *J. Comput. Chem.* **2010**, *31*, 224–247.

(25) Roos, B. O.; Lindh, R.; Malmqvist, P.-Å.; Veryazov, V.; Widmark, P.-O. *J. Phys. Chem. A* **2004**, *108*, 2851–2858.

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

(27) Douglas, N.; Kroll, N. M. *Ann. Phys. (N. Y.)* **1974**, *82*, 89.

(28) Hess, B. *Phys. Rev. A* **1986**, *33*, 3742.

(29) Roos, B. O.; Malmqvist, P.-Å. *Phys. Chem. Chem. Phys.* **2004**, *6*, 2919–2927.

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

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

(32) Forsberg, N.; Malmqvist, P.-Å. *Chem. Phys. Lett.* **1997**, *274*, 196.

(33) Aquilante, F.; Pedersen, T. B.; Lindh, R. *J. Chem. Phys.* **2007**, *126*, 194106.

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

(35) Aquilante, F.; Pedersen, T. B.; Lindh, R. *Theor. Chem. Acc.* **2009**, *124*, 1–10.

(36) Marsden, C. J.; Moncrieff, D.; Quelch, G. E. *J. Phys. Chem.* **1994**, *98*, 2038.

(37) Vanquickenborne, L. G.; Vinckier, A. E.; Pierloot, K. *Inorg. Chem.* **1996**, *35*, 1305–1309.

(38) Ryan, M. F.; Eyley, J. R.; Richardson, D. E. *J. Am. Chem. Soc.* **1992**, *114*, 8611–8619.

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

(40) Klopper, W.; Luthi, H. P. *Chem. Phys. Lett.* **1996**, *262*, 546–552.

(41) Koch, H.; Jorgensen, P.; Helgaker, T. *J. Chem. Phys.* **1996**, *104*, 9528–9530.

(42) Haaland, A. *Top. Curr. Chem.* **1975**, *53*, 1.

(43) Beach, N. A.; Gray, H. B. *J. Am. Chem. Soc.* **1968**, *90*, 5713.

(44) Pollak, C.; Rosa, A.; Baerends, E. J. *J. Am. Chem. Soc.* **1997**, *119*, 7324–7329.

(45) Rosa, A.; Baerends, E.; van Gisbergen, S. J. A.; van Lenthe, A.; Groeneveld, J. A.; Snijders, J. G. *J. Am. Chem. Soc.* **1999**, *121*, 10356.

(46) Hummel, P.; Oxgaard, J.; Goddard, W. A., III; Gray, H. B. *Inorg. Chem.* **2005**, *44*, 2454–2458.

(47) Crespo-Otero, R.; Barbatti, M. *J. Chem. Phys.* **2011**, *134*, 164305.

(48) Ben Amor, N.; Villaume, S.; Maynau, D.; Daniel, C. *Chem. Phys. Lett.* **2006**, *421*, 378–382.

(49) Villaume, S.; Strich, A.; Daniel, C.; Perera, S. A.; Bartlett, R. J. *Phys. Chem. Chem. Phys.* **2007**, *9*, 6115–6122.

(50) Rees, B.; Mitschler, A. *J. Am. Chem. Soc.* **1976**, *98*, 7918.

(51) Finley, J.; Malmqvist, P.-Å.; Roos, B. O.; Serrano-Andrés, L. *Chem. Phys. Lett.* **1998**, *288*, 299–306.

(52) Collman, J. P.; Hoard, J. L.; Kim, N.; Lang, G.; Reed, C. A. *J. Am. Chem. Soc.* **1975**, *97*, 2676–2681.

(53) Goff, H.; La Mar, G. N.; Reed, C. A. *J. Am. Chem. Soc.* **1977**, *99*, 3641–3646.

(54) Kitagawa, T.; Teraoka, J. *Chem. Phys. Lett.* **1979**, *63*, 443–446.

(55) Dolphin, D.; Sams, J. R.; Tsin, T. B.; Wong, K. L. *J. Am. Chem. Soc.* **1976**, *98*, 6970–6975.

(56) Lang, G.; Spertalian, K.; Reed, C. A.; Collman, J. P. *J. Chem. Phys.* **1978**, *69*, 5424–5427.

(57) Boyd, P. D. W.; Buckingham, A. D.; McMeeking, R. F.; Mitra, S. *Inorg. Chem.* **1979**, *18*, 3585–3591.

(58) Mispelter, J.; Momenteau, M.; Lhoste, J. M. *J. Chem. Phys.* **1980**, *72*, 1003–1012.

(59) Strauss, S. H.; Silver, M. E.; Long, K. M.; Thompson, R. G.; Hudgens, R. A.; Spertalian, K.; Ibers, J. A. *J. Am. Chem. Soc.* **1985**, *107*, 4207–4215.

(60) *Iron Porphyrin*; Lever, A. B. P., Gray, H. B., Eds. Addison-Wesley, Inc.: Reading, MA, 1983.

(61) Obara, S.; Kashiwagi, H. *J. Chem. Phys.* **1982**, *77*, 3155.

(62) Dedieu, A.; Rohmer, M.-M.; Veillard, A. *Adv. Quantum Chem.* **1982**, *16*, 43–95.

(63) Rawlings, D. C.; Gouterman, M.; Davidson, E.; Feller, D. *Int. J. Quantum Chem.* **1985**, *28*, 773–796.

(64) Kozłowski, P. M.; Spiro, T. G.; Bérces, A.; Zgierski, M. Z. *J. Phys. Chem. B* **1998**, *102*, 2603–2608.

(65) Choe, Y.-K.; Hashimoto, T.; Nakano, H.; Hirao, K. *Chem. Phys. Lett.* **1998**, *295*, 380–388.

(66) Choe, Y.-K.; Nakajima, T.; Hirao, K.; Lindh, R. *J. Chem. Phys.* **1999**, *111*, 3837–3844.

(67) Scherlis, D. A.; Estrin, D. A. *Int. J. Quantum Chem.* **2002**, *87*, 158–166.

(68) Liao, M.-S.; Scheiner, S. *J. Chem. Phys.* **2002**, *117*, 205–219.

(69) Deeth, R. *J. Struct. Bonding (Berlin, Ger.)* **2004**, *113*, 37–69.

(70) Deeth, R. J.; Fey, N. *J. Comput. Chem.* **2004**, *25*, 1840–1848.

(71) Groenhof, A. R.; Swart, M.; Ehlers, A. E.; Lammertsma, K. *J. Phys. Chem. A* **2005**, *109*, 3411–3417.

(72) Liao, M.-S.; Watts, J. D.; Huang, M.-J. *J. Phys. Chem. A* **2007**, *111*, 5927–5935.

(73) Khvostichenko, D.; Choi, A.; Boulatov, R. *J. Phys. Chem. A* **2008**, *112*, 3700–3711.

(74) Ahlrichs, R.; Bär, M.; Häser, M.; Horn, H.; Kölmel, C. *Chem. Phys. Lett.* **1989**, *162*, 165–169.

(75) Treutler, O.; Ahlrichs, R. *J. Chem. Phys.* **1995**, *102*, 346.

(76) Pierloot, K.; Tsokos, E.; Roos, B. O. *Chem. Phys. Lett.* **1993**, *214*, 583.

(77) Malmqvist, P.-Å.; Roos, B. O.; Schimmelpfennig, B. *Chem. Phys. Lett.* **2002**, *357*, 230–240.

(78) Hess, B. A.; Marian, C. M.; Wahlgren, U.; Gropen, O. *Chem. Phys. Lett.* **1996**, *251*, 365–371.

(79) Christiansen, O.; Gauss, J.; Schimmelpfennig, B. *Phys. Chem. Chem. Phys.* **2000**, *2*, 965–971.

(80) Vahtras, O.; Engström, M.; Schimmelpfennig, B. *Chem. Phys. Lett.* **2002**, *351*, 424–430.

(81) Vancoillie, S.; Rulíšek, L.; Neese, F.; Pierloot, K. *J. Phys. Chem. A* **2009**, *113*, 6149–6157.

(82) Ghosh, A.; Wondimagegn, T.; Parusel, A. B. *J. Am. Chem. Soc.* **2000**, *122*, 5100–5104.

(83) Bröring, M.; Brégier, F.; Cónsul Tejero, E.; Hell, C.; Holthausen, M. C. *Angew. Chem., Int. Ed.* **2007**, *46*, 445–448.

(84) Paolesse, R.; Kadish, K. M.; Smith, K. M.; Guillard, R. *The Porphyrin Handbook*; Academic Press: Boston, MA, 2000; Vol. 2; pp 201–300.

(85) Paolesse, R. *SYNLETT* **2008**, *15*, 2215–2230.

(86) Grodkowski, J.; Neta, P.; Fujita, E.; Mahammed, A.; Simkhovich, L.; Gross, Z. *J. Phys. Chem. A* **2002**, *106*, 4772–4778.

- (87) Will, S.; Lex, J.; Vogel, E.; Adamian, V. A.; Caemelbecke, E. V.; Kadish, K. M. *Inorg. Chem.* **1996**, *35*, 5577–5583.
- (88) Guillard, R.; Gros, C. P.; Bolze, F.; Jérôme, F.; Ou, Z. P.; Shao, J. G.; Fischer, J.; Weiss, R.; Kadish, K. M. *Inorg. Chem.* **2001**, *40*, 4845.
- (89) Kadish, K. M.; Ou, Z. P.; Shao, J. G.; Gros, C. P.; Barbe, J. M.; Jérôme, F.; Bolze, F.; Burdet, F.; Guillard, R. *Inorg. Chem.* **2002**, *41*, 3990.
- (90) Maiti, N.; Lee, J.; Kwon, S. J.; Kwak, J.; Do, Y.; Churchill, D. G. *Polyhedron* **2006**, *25*, 1519–1530.
- (91) Barbe, J. M.; Canard, G.; Brandés, S.; Guillard, R. *Angew. Chem.* **2005**, *117*, 3163–3166.
- (92) Guillard, R.; Burdet, F.; Barbe, J. M.; Gros, C. P.; Espinosa, E.; Shao, J.; Ou, R.; Kadish, K. M. *Inorg. Chem.* **2005**, *44*, 3972–3983.
- (93) Rovira, C.; Kunc, K.; Hutter, J.; Parrinello, M. *Inorg. Chem.* **2001**, *40*, 11–17.
- (94) Andersen, T.; Lykke, K. R.; Neumark, D. M.; Lineberger, W. C. *J. Chem. Phys.* **1987**, *86*, 1858.
- (95) Engelking, P. C.; Lineberger, W. C. *J. Chem. Phys.* **1977**, *66*, 5054.
- (96) Cheung, A. S.-C.; Lee, N.; Lyrre, A. M.; Merer, A. J.; Taylor, A. W. *J. Mol. Spectrosc.* **1982**, *95*, 213.
- (97) Fan, J.; Wang, L.-S. *J. Chem. Phys.* **1995**, *102*, 8714.
- (98) Wu, H. B.; Desai, S. R.; Wang, L. S. *J. Am. Chem. Soc.* **1996**, *118*, 5296.
- (99) Drechsler, G.; Boesl, U.; Bassmann, C.; Schlag, E. W. *J. Chem. Phys.* **1997**, *107*, 2284.
- (100) Neumark, D. M.; Lineberger, W. C. *J. Phys. Chem. A* **2009**, *113*, 10588.
- (101) Sakellaris, C. N.; Milliordos, E.; Mavridis, A. *J. Chem. Phys.* **2011**, *134*, 234308.

Dispersion-Weighted Explicitly Correlated Coupled-Cluster Theory [DW-CCSD(T^{**})-F12]

Michael S. Marshall and C. David Sherrill*

Center for Computational Molecular Science and Technology, School of Chemistry and Biochemistry, Georgia Institute of Technology, Atlanta, Georgia 30332-0400, United States

ABSTRACT: We propose a procedure denoted dispersion-weighted explicitly correlated coupled-cluster [DW-CCSD(T^{**})-F12] which mixes CCSD(T^{**})-F12a and CCSD(T^{**})-F12b so as to correct the small errors exhibited by each of the approximations in a small basis set, allowing for a black-box method that can provide high-quality interaction energies for a variety of nonbonded interactions. Relative to CCSD(T^{**})-F12a and CCSD(T^{**})-F12b, DW-CCSD(T^{**})-F12 reduces the mean absolute deviation by a factor of 2 and the maximum error by a factor of 3 (formic acid dimer) and 4 (stacked adenine–thymine) for the aug-cc-pVDZ basis set.

1. INTRODUCTION

Coupled-cluster with single, double, and perturbative triple excitations [CCSD(T)] has been referred to as the “gold standard” in computational chemistry.¹ This sophisticated description of dynamic electron correlation is often required to achieve benchmark quality results. In the context of noncovalent interactions, CCSD(T) can provide very accurate results,² but this typically requires very large basis sets, augmented with diffuse functions. Coupled with the steep computational scaling of CCSD(T), this significantly restricts the size of systems that may be studied at this level of accuracy. With the introduction of explicitly correlated wave functions, however, this computational cost is severely abated because accurate energies may be attained using relatively small basis sets.^{3–14} Applying these methods to noncovalent interactions, Marchetti et al.¹⁵ showed that CCSD(T^{**})-F12a/aug-cc-pVDZ can achieve an accuracy of better than 0.2 kcal mol⁻¹ for all dimers in the S22 test set.¹⁶ They argued that this same accuracy would require at least an aug-cc-pVQZ basis set with traditional CCSD(T), which would be 1–2 orders of magnitude more expensive. More recently, de Lange et al.¹⁴ demonstrated that CCSD(T)-F12a/VDZ-F12 and CCSD(T)-F12b/VDZ-F12 achieve an average accuracy of 0.03 kcal mol⁻¹ compared to CCSD(T)/CBS(a5Z,a6Z) for small molecules interacting with carbon dioxide. This accuracy is impressive considering the standard CCSD(T)/aug-cc-pVDZ has an average error of 0.4 kcal mol⁻¹ for these complexes.

When employing explicitly correlated methods, the choice of ansatz and how to treat the triples correction are both very important. For coupled-cluster, the F12a and F12b approximations have become the most commonly used.^{9,12} Within the MOLPRO package¹⁷ used here, since there is no explicitly correlated triples correction, this quantity should be scaled to achieve highly accurate results. This work follows the approach of Werner and co-workers^{12,15} whereby the triples correction is scaled by the ratio of MP2 correlation energy and MP2-F12 correlation energy:

$$E_{(T^*)}^{\text{corr}} = E_{(T)}^{\text{corr}} \cdot \frac{E_{\text{MP2-F12}}}{E_{\text{MP2}}} \quad (1)$$

To retain size-consistency for interaction energies, one must use the same scale factor for all computations (the dimer, and both monomers). Because of this difference, we designate CCSD(T^{**})-F12 (2 asterisks) to refer to the size-consistent version (using the dimer scale factor for all three computations) and CCSD(T^{*})-F12 (1 asterisk) to refer to the independently scaled version.

In our recent study,¹⁸ CCSD(T^{**})-F12a/aug-cc-pVDZ is shown to yield a mean absolute deviation (MAD) of 0.09 kcal mol⁻¹ against the newly revised interaction energies for the S22 test set¹⁶ (herein referred to as S22B).¹⁸ This accuracy is notable because standard CCSD(T) (in the absence of focal-point analysis) would require a much larger basis set, such as aug-cc-pVQZ, to reach this accuracy, and such computations would become prohibitively expensive for all but the smallest molecular systems. Notably, CCSD(T^{**})-F12a reproduces benchmark energies for hydrogen-bonded systems very accurately, while incurring small errors for dispersion-bound complexes, and CCSD(T^{**})-F12b reproduces benchmark energies for dispersion-bound complexes very accurately but degrades somewhat in quality for hydrogen bonding. A very similar problem has been investigated by Marchetti et al.¹⁵ in their studies of explicitly correlated spin-component-scaled Møller–Plesset perturbation theory (SCS-MP2-F12) and MP2-F12. In their work, they found MP2 significantly overbinds dispersion-bound complexes but does well for hydrogen bonding, while SCS-MP2-F12 can properly describe dispersion but underbinds hydrogen-bonding complexes. They proposed mixing the two approaches in a method known as dispersion-weighted MP2 (DW-MP2):

$$\Delta E_{\text{DW-MP2}} = \omega \Delta E_{\text{MP2-F12}} + (1 - \omega) \Delta E_{\text{SCS-MP2-F12}} \quad (2)$$

$$\omega = \frac{1}{2} \left[1 + \tanh \left(\alpha + \beta \frac{\Delta E_{\text{SCF}}}{\Delta E_{\text{MP2-F12}}} \right) \right] \quad (3)$$

The chosen switching function (eq 3) is a hyperbolic tangent function with two fit parameters, and the switching metric between SCS-MP2-F12 and a MP2-F12 is the ratio of MP2-F12 and

Received: August 25, 2011

Published: October 17, 2011

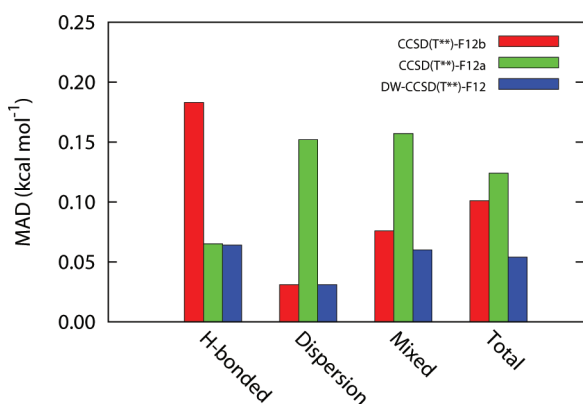


Figure 1. MAD of interaction energies versus S22B benchmark CCSD(T)/CBS values (ref 18) for explicitly correlated methods and DW-CCSD(T**)–F12 using the aug-cc-pVDZ basis.

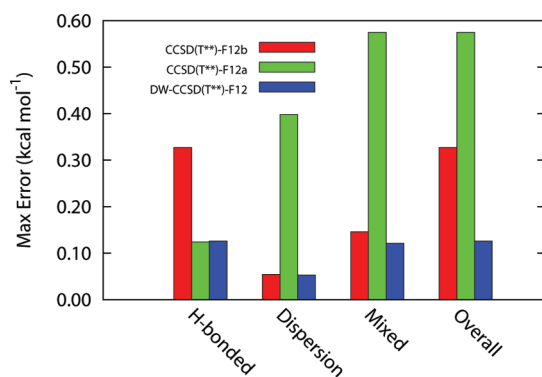


Figure 2. MAD of interaction energies versus S22B benchmark CCSD(T)/CBS (ref 18) values for explicitly correlated methods and DW-CCSD(T**)–F12 using the aug-cc-pVDZ basis.

self-consistent field (SCF) interaction energies. The underlying concept is that hydrogen-bonded complexes have a ratio near one because the interaction is predominantly electrostatic, which is properly captured by SCF, while dispersion-dominated systems should yield a ratio far from one, as SCF fails to model dispersion because of its lack of dynamic electron correlation. This mixing transforms MP2-F12 [root-mean-square error (RMSE) of 1.17 kcal mol^{−1} for the aug-cc-pVDZ basis] and SCS-MP2-F12 (1.08 kcal mol^{−1}) into DW-MP2 (0.24 kcal mol^{−1}) for the S22B test set. Inspired by this procedure, we explore analogous methods of combining CCSD(T**)–F12a and CCSD(T**)–F12b based on the character of the noncovalent interaction.

2. THEORETICAL METHODS

As demonstrated in Figures 1 and 2, direct computation of interaction energies by CCSD(T**)–F12, even when using the modest aug-cc-pVDZ basis set, yields surprisingly accurate values compared to the best available estimates. Here, we seek to correct the minor remaining deficiencies in explicitly correlated CCSD(T**)–F12a and CCSD(T**)–F12b with a scheme similar to the DW-MP2 approach of Marchetti et al.¹⁵

Using eqs 3 and 4, we fit α and β against the S22B test set. The SCF energy in eq 3 includes the complementary auxiliary orbital basis (CABS) correction.^{19,20} To test the transferability of these

fit parameters, we also consider another test set, HSG-A.^{18,21}

$$\Delta E_{\text{DW-CCSD(T**)–F12}} = \omega \Delta E_{\text{CCSD(T**)–F12a}} + (1 - \omega) \Delta E_{\text{CCSD(T**)–F12b}} \quad (4)$$

The two test sets considered here, S22 and HSG, were chosen for their small size, convenient separation into hydrogen-bonded, dispersion-dominated, and mixed bonding classes and quality of available benchmark interaction energies. S22 is an established test set with systems ranging from water dimer to adenine–thymine complexes. Its reference values have recently been revised (S22B) to be of MP2/CBS(aug-cc-pVTZ, aug-cc-pVQZ) + $\delta_{\text{MP2}}^{\text{CCSD(T)}}$ /aug-cc-pVTZ quality or better. For the 10 smallest complexes in the S22 test set, at least MP2/CBS(aug-cc-pVQZ, aug-cc-pV5Z) + $\delta_{\text{MP2}}^{\text{CCSD(T)}}$ /aug-cc-pVQZ quality is achieved to ensure that we can reliably benchmark both DW-CCSD(T**)–F12/aug-cc-pVDZ (against the entire S22 test set) and DW-CCSD(T**)–F12/aug-cc-pVTZ (against the ten smallest complexes in the S22 test set). The HSG test set was formed by dissecting the binding site of a bound protein–drug complex (HIV-II protease/indinavir) into 21 pairs of chemical fragments (each of which is not necessarily at its individual equilibrium geometry).²¹ We recently revised¹⁸ the HSG benchmark energies replacing the previous extrapolated heavy-aug-cc-pVDZ/heavy-aug-cc-pVTZ $\delta_{\text{MP2}}^{\text{CCSD(T)}}$ corrections with the bare heavy-aug-cc-pVTZ $\delta_{\text{MP2}}^{\text{CCSD(T)}}$ corrections to avoid overestimating the coupled-cluster correction for hydrogen-bonded systems [here heavy-aug-cc-pVXZ refers to the aug-cc-pVXZ basis where diffuse functions are added only to heavy (nonhydrogen) atoms]. This revision results from the recent systematic study of the nonmonotonic convergence of $\delta_{\text{MP2}}^{\text{CCSD(T)}}$ for various members of the S22 test set.¹⁸ The revised HSG-A values are thus similar in accuracy to the S22B values. All interaction energies were counterpoise corrected for basis set superposition error (BSSE) using the scheme outlined by Boys and Bernardi.²²

3. RESULTS AND DISCUSSION

Using eqs 3 and 4, we smoothly combine the F12a to F12b ansätze according to the relative difference in ΔE_{SCF} and $\Delta E_{\text{MP2-F12}}$ interaction energies. Figure 1 and Table 1 illustrate the errors associated with each ansatz independently, and it is clear that some combination of the two should result in a method that captures both types of interactions accurately on average. We find fitted parameters of $\alpha = -1$ and $\beta = 4$ for the S22B set, which results in a MAD of 0.06, 0.03, 0.07, and 0.05 kcal mol^{−1}, respectively, for hydrogen-bonding, dispersion-dominated, and mixed interactions and overall. The complexes were assigned to these categories according to the SAPT2+(3)/aug-cc-pVTZ results of Hohenstein and Sherrill.²³ The most substantial gain of this approach is in the reduction of maximum errors for each subset, as shown in Figure 2. The maximum errors for F12a and F12b are 0.58 and 0.33 kcal mol^{−1}, respectively, whereas the largest error incurred by DW-CCSD(T**)–F12 is 0.13 kcal mol^{−1} for formamide dimer. This is a factor of 3–4 reduction in the maximum error and a factor of 2 reduction in the MAD. Figure 3 shows the error for each complex in the S22B test set for CCSD(T**)–F12a, CCSD(T**)–F12b, and DW-CCSD(T**)–F12. In this figure, one can clearly see how DW-CCSD(T**)–F12 switches between CCSD(T**)–F12a and CCSD(T**)–F12b as appropriate to avoid the largest maximum errors.

Table 1. Counterpoise Corrected Interaction Energies (kcal mol⁻¹) for the CCSD(T^{})-F12a, CCSD(T^{**})-F12b, and DW-CCSD(T^{**})-F12 Methods Compared to S22B Benchmark Values (ref 18)^a**

	complex	reference IE	F12a	F12b	DW-F12
1	HB ammonia dimer, C _{2h}	-3.133	-3.11	-3.05	-3.10
2	HB water dimer, C _s	-4.989	-4.92	-4.86	-4.92
3	HB formic acid dimer, C _{2h}	-18.753	-18.63	-18.43	-18.63
4	HB formamide dimer, C _{2h}	-16.062	-15.94	-15.80	-15.94
5	HB hydrogen-bonded uracil dimer, C _{2h}	-20.641	-20.63	-20.45	-20.63
6	HB 2-pyridone-2-aminopyridine, C ₁	-16.934	-16.98	-16.79	-16.97
7	HB adenine-thymine WC, C ₁	-16.660	-16.72	-16.51	-16.71
8	DD methane dimer, D _{3d}	-0.527	-0.53	-0.51	-0.51
9	DD ethene dimer, D _{2d}	-1.472	-1.50	-1.44	-1.44
10	DD benzene-methane, C ₃	-1.448	-1.47	-1.40	-1.40
11	DD parallel displaced benzene dimer, C _{2h}	-2.654	-2.90	-2.64	-2.64
12	DD pyrazine dimer, C _s	-4.255	-4.54	-4.26	-4.26
13	MX stacked uracil dimer, C ₂	-9.805	-10.17	-9.84	-9.88
14	DD stacked indole-benzene, C ₁	-4.524	-4.92	-4.57	-4.57
15	MX stacked adenine-thymine, C ₁	-11.730	-12.30	-11.84	-11.85
16	MX ethene-ethine, C _{2v}	-1.496	-1.51	-1.47	-1.49
17	MX benzene-water, C _s	-3.275	-3.23	-3.13	-3.19
18	MX benzene-ammonia, C _s	-2.312	-2.31	-2.23	-2.23
19	MX benzene-hydrogen cyanide, C _s	-4.541	-4.49	-4.39	-4.47
20	DD T-shaped benzene dimer, C _{2v}	-2.717	-2.78	-2.66	-2.66
21	MX T-shaped indole-benzene, C ₁	-5.627	-5.73	-5.56	-5.58
22	MX phenol dimer, C ₁	-7.097	-7.15	-6.99	-7.07
		Hydrogen Bonded			
	maximal deviation		0.12	0.33	0.13
	mean signed deviation		0.03	0.18	0.04
	MAD		0.06	0.18	0.06
	RMSD		0.08	0.20	0.08
		Mixed Influence			
	Maximal Deviation		-0.57	0.15	-0.12
	Mean Signed Deviation		-0.13	0.05	0.02
	MAD		0.15	0.09	0.07
	RMSD		0.25	0.10	0.07
		Dispersion Bound			
	Maximal Deviation		-0.01	0.05	0.05
	Mean Signed Deviation		-0.40	0.02	0.02
	MAD		0.15	0.03	0.03
	RMSD		0.21	0.04	0.04
		Full Set			
	Maximal Deviation		-0.57	0.33	0.13
	Mean Signed Deviation		-0.08	0.08	0.02
	MAD		0.12	0.10	0.05
	RMSD		0.19	0.13	0.07

^a Fit parameters are $\alpha = -1$ and $\beta = 4$. Computations use the aug-cc-pVDZ basis set. The errors and mixing in the DW approach can be seen graphically in Figure 3.

To evaluate the transferability of fitting parameters, DW-CCSD(T^{**})-F12/aug-cc-pVDZ was applied to the HSG-A test set using the parameters fit against the S22B test set; it achieves a MAD of 0.04 kcal mol⁻¹ for overall interaction energies. This is a significant achievement, especially because the diverse, nonequilibrium nature of the complexes found in the HSG-A set is often challenging for computational methods. The maximum

error across this test set is 0.10 kcal mol⁻¹. These impressive results compare to a MAD of 0.42 kcal mol⁻¹ and a maximum error of 1.53 kcal mol⁻¹ for conventional CCSD(T)/heavy-aug-cc-pVDZ.

DW-CCSD(T^{**})-F12 was also tested using an aug-cc-pVTZ basis set against the 10 smallest complexes in the S22B benchmark set Figure 4 and Table 2. With fit parameters $\alpha = 0.4$ and

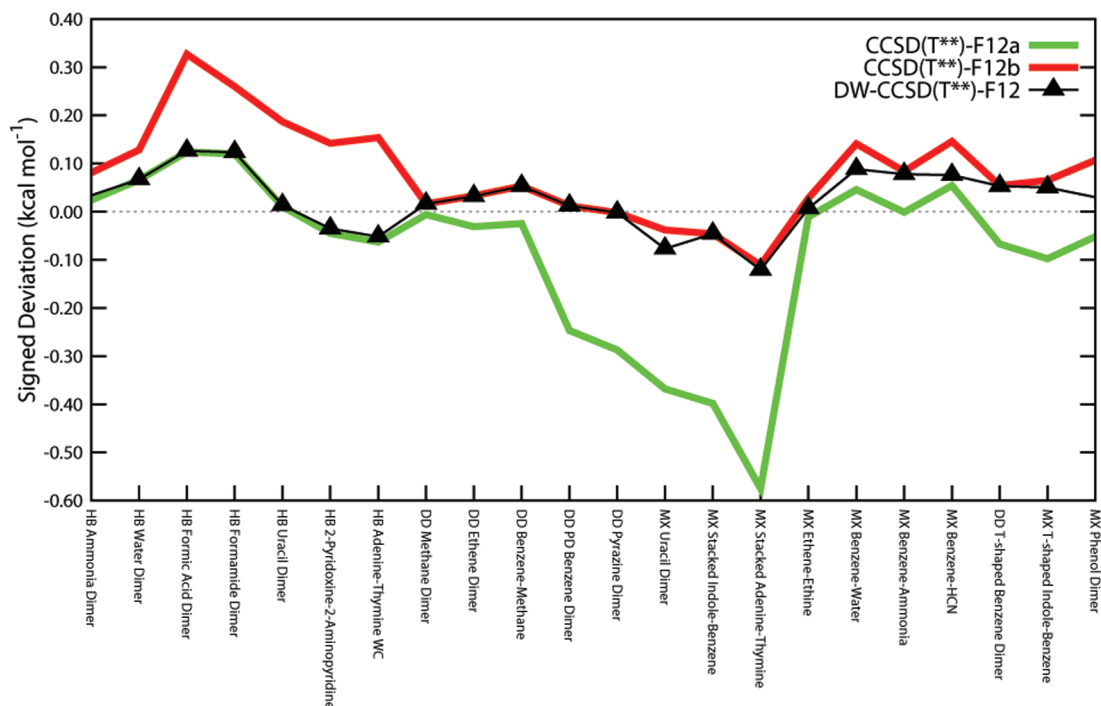


Figure 3. For the aug-cc-pVDZ basis set, interaction energy errors for CCSD(T^{**})-F12a, CCSD(T^{**})-F12b, and DW-CCSD(T^{**})-F12 methods for each complex in the S22 test set. All errors in kcal mol⁻¹, relative to S22B CCSD(T)/CBS benchmarks (ref 18). Individual errors and statistics can be found in Table 1.

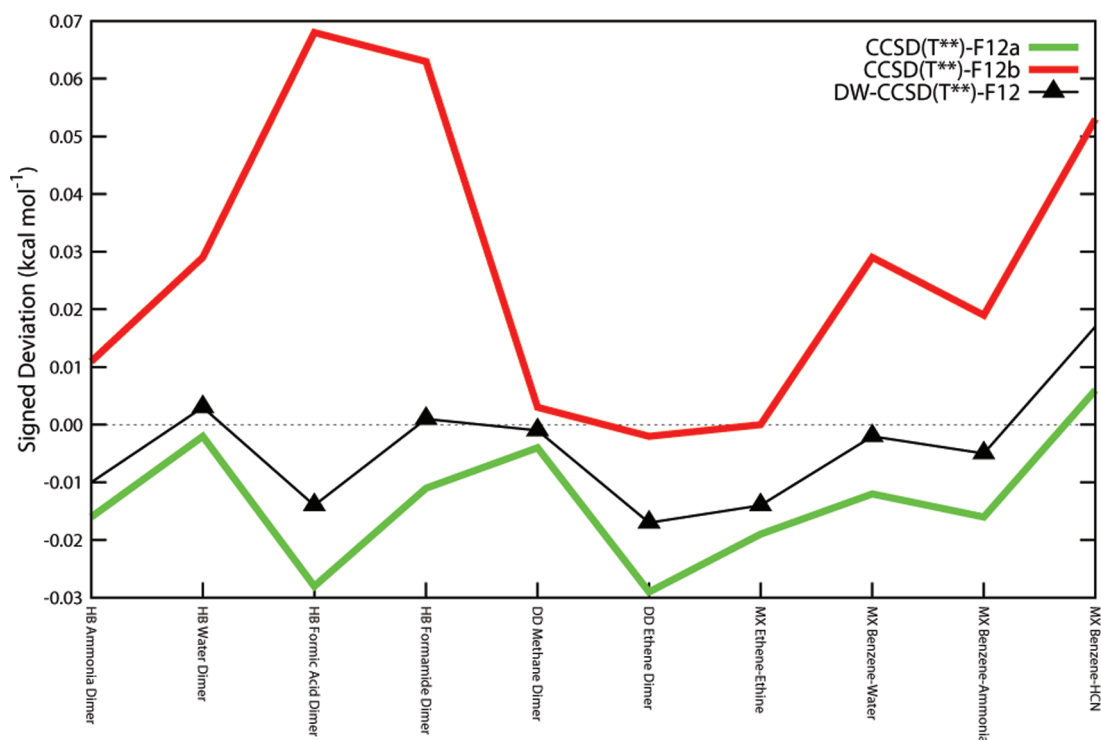


Figure 4. For the aug-cc-pVTZ basis set, interaction energy errors for CCSD(T^{**})-F12a, CCSD(T^{**})-F12b, and DW-CCSD(T^{**})-F12 methods for each complex in the S22 test set. All errors in kcal mol⁻¹, relative to a subset of the S22B CCSD(T)/CBS benchmarks (ref 18).

$\beta = 0.6$, we find MADs of 0.028, 0.014, and 0.008 kcal mol⁻¹ and maximum errors of 0.029, 0.068, and 0.017 kcal mol⁻¹ for CCSD(T^{**})-F12a, CCSD(T^{**})-F12b, and DW-CCSD(T^{**})-F12,

respectively. When using the aug-cc-pVDZ fit parameters of $\alpha = -1$ and $\beta = 4$, DW-CCSD(T^{**})-F12 still achieves an overall MAD of 0.011 kcal mol⁻¹ and a maximum error of 0.027 kcal mol⁻¹

Table 2. Counterpoise Corrected Interaction Energies (kcal mol⁻¹) for the CCSD(T^{})-F12a, CCSD(T^{**})-F12b, and DW-CCSD(T^{**})-F12 Methods Compared to S22B Benchmark Values (ref 18)^a**

complex	reference IE	F12a	F12b	DW-F12
1 HB ammonia dimer, C _{2h}	-3.133	-3.15	-3.12	-3.14
2 HB water dimer, C _s	-4.989	-4.99	-4.96	-4.99
3 HB formic acid dimer, C _{2h}	-18.753	-18.78	-18.69	-18.77
4 HB formamide dimer, C _{2h}	-16.062	-16.07	-16.00	-16.06
8 DD methane dimer, D _{3d}	-0.527	-0.53	-0.52	-0.53
9 DD ethene dimer, D _{2d}	-1.472	-1.50	-1.47	-1.49
16 MX ethene-ethine, C _{2v}	-1.496	-1.52	-1.50	-1.51
17 MX benzene-water, C _s	-3.275	-3.29	-3.25	-3.28
18 MX benzene-ammonia, C _s	-2.312	-2.33	-2.29	-2.32
19 MX benzene-hydrogen cyanide, C _s	-4.541	-4.54	-4.49	-4.53
maximal deviation		0.029	0.068	0.017
MAD		0.014	0.028	0.008

^a Fit parameters are $\alpha = 0.4$ and $\beta = 0.6$. Computations use the aug-cc-pVTZ basis set. The errors and mixing in the DW approach can be seen graphically in Figure 4.

(formic acid dimer). DW-CCSD(T^{**})-F12 was also tested using the heavy-aug-cc-pVDZ basis set, which is aug-cc-pVDZ on the nonhydrogen atoms and cc-pVDZ on the hydrogens. DW-CCSD(T^{**})-F12/heavy-aug-cc-pVDZ achieves an MAD of 0.131 kcal mol⁻¹ for the S10 subset (compared to 0.065 kcal mol⁻¹ for this subset when using the aug-cc-pVDZ basis).

4. CONCLUSION

A new “dispersion-weighted” approach is proposed for explicitly correlated coupled-cluster studies of weakly bound systems that accurately describes both hydrogen-bonding and dispersion-bound complexes through a black-box admixture of CCSD(T^{**})-F12a and CCSD(T^{**})-F12b. This technique achieves MAD values of 0.05 and 0.04 kcal mol⁻¹ for the S22B and HSG-A test sets (using S22B-fit parameters), corresponding to a factor of 2 reduction of the MAD and a factor of 3 and 4 reduction in maximum error relative to F12a and F12b methods, respectively. Explicitly correlated CCSD-F12 computations are more expensive than canonical CCSD, but the extra cost is usually negligible with the inclusion of perturbative triples. This suggests that CCSD(T^{**})-F12, and perhaps the dispersion-weighted variant proposed here, should become a preferred approach for obtaining accurate benchmarks for noncovalent interactions.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: sherrill@gatech.edu.

■ ACKNOWLEDGMENT

This material is based upon work supported by the National Science Foundation (grant no. CHE-1011360). M.S.M. would like to thank Dr. Lori Burns for valuable discussions and analysis. The computer resources of the Center for Computational Molecular Science and Technology are funded through a National Science Foundation CRIF Award (CHE-0946869).

■ REFERENCES

- Lee, T. J.; Scuseria, G. E. In *Quantum Mechanical Electronic Structure Calculations with Chemical Accuracy*; Langhoff, S. R., Ed.; Kluwer Academic Publishers: Dordrecht, The Netherlands, 1995; pp 47–108.
- Hopkins, B. W.; Tschumper, G. S. *J. Phys. Chem. A* **2004**, *108*, 2941–2948.
- May, A. J.; Valeev, E. F.; Polly, R.; Manby, F. R. *Phys. Chem. Chem. Phys.* **2005**, *7*, 2710.
- Klopper, W.; Manby, F. R.; Ten-no, S.; Valeev, E. F. *Int. Rev. Phys. Chem.* **2006**, *25*, 427–468.
- Werner, H.-J.; Manby, F. R. *J. Chem. Phys.* **2006**, *124*, 054114.
- Manby, F. R.; Werner, H.-J.; Adler, T. B.; May, A. J. *J. Chem. Phys.* **2006**, *124*, 094103.
- Tew, D. P.; Klopper, W.; Neiss, C.; Hättig, C. *Phys. Chem. Chem. Phys.* **2007**, *9*, 1921.
- Werner, H.-J.; Adler, T. B.; Manby, F. R. *J. Chem. Phys.* **2007**, *126*, 164102.
- Adler, T. B.; Knizia, G.; Werner, H.-J. *J. Chem. Phys.* **2007**, *127*, 221106.
- Werner, H.-J. *J. Chem. Phys.* **2008**, *124*, 101103.
- Klopper, W.; Ruscic, B.; Tew, D. P.; Bischoff, F. A.; Wolfsegger, S. *Chem. Phys.* **2009**, *356*, 14.
- Knizia, G.; Adler, T.; Werner, H. J. *J. Chem. Phys.* **2009**, *130*, 054104.
- Shiozaki, T.; Kamiya, M.; Hirata, S.; Valeev, E. F. *J. Chem. Phys.* **2009**, *130*, 054101.
- de Lange, K. M.; Lane, J. R. *J. Chem. Phys.* **2011**, *134*, 034301.
- Marchetti, O.; Werner, H. J. *Phys. Chem. A* **2009**, *113*, 11580–11585.
- Jurečka, P.; Šponer, J.; Černý, J.; Hobza, P. *Phys. Chem. Chem. Phys.* **2006**, *8*, 1985–1993.
- Werner, H.-J.; Knowles, P. J.; Knizia, G.; Manby, F. R.; Schütz, M.; Celani, P.; Korona, T.; Lindh, R.; Mitrushenkov, A.; Rauhut, G.; Shamasundar, K. R.; 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.; Hesselmann, A.; Hetzer, G.; Hrenar, T.; Jansen, G.; Köppl, C.; Liu, Y.; Lloyd, A. W.; Mata, R. A.; May, A. J.; McNicholas, S. J.; Meyer, W.; Mura, M. E.; Nicklass, A.; O'Neill, D. P.; Palmieri, P.; Pflüger, K.; Pitzer, R.; Reiher, M.; Shiozaki, T.; Stoll, H.; Stone, A. J.; Tarroni, R.; Thorsteinsson, T.; Wang, M.; Wolf, A. *MOLPRO*, version 2010.1; University College Cardiff Consultants Limited: Cardiff, U.K.; see <http://www.molpro.net>.
- Marshall, M. S.; Burns, L. A.; Sherrill, C. D. *J. Chem. Phys.*, in press.
- Valeev, E. F. *Chem. Phys. Lett.* **2004**, *395*, 190–195.
- Knizia, G.; Werner, H.-J. *J. Chem. Phys.* **2008**, *128*, 154103.
- Faver, J. C.; Benson, M. L.; He, X. A.; Roberts, B. P.; Wang, B.; Marshall, M. S.; Kennedy, M. R.; Sherrill, C. D.; Merz, K. M. *J. Chem. Theory Comput.* **2011**, *7*, 790.
- Boys, S. F.; Bernardi, F. *Mol. Phys.* **1970**, *19*, 553–566.
- Hohenstein, E. G.; Sherrill, C. D. *WIREs Comput. Mol. Sci.*, in press.

Spline Implementation of Generalized Gradient Approximations to the Exchange–Correlation Functional and Study of the Sensitivity of Density Functional Accuracy to Localized Domains of the Reduced Density Gradient

Roberto Peverati and Donald G. Truhlar*

Department of Chemistry and Supercomputing Institute, University of Minnesota, Minneapolis, Minnesota 55455, United States

ABSTRACT: We present a natural cubic spline implementation of the exchange enhancement factor as a function of the reduced density gradient, and we demonstrate its performance by replicating the results of common GGA functionals. We also investigate the effect on the accuracy of various calculated properties of changing the shape of the exchange enhancement factor and an analogous factor for correlation. The properties considered are main group atomization energies, ionization potentials, electron affinities, proton affinities, alkyl bond dissociation energies, difficult hydrocarbon cases, barrier heights for chemical reactions, noncovalent interactions, atomic energies, metal bond energies, and main group bond lengths.

1. INTRODUCTION

The success of Kohn–Sham density functional theory rests on the accuracy with which one approximates the exchange–correlation functional, usually just called the density functional.¹ Although modern density functionals often depend on a number of variables including the density, the density gradient, the Laplacian of the density, the orbital-dependent Hartree–Fock energy, and the orbital-dependent kinetic energy density, they are almost all built by adding ingredients to the generalized gradient approximation (GGA), in which the density functional depends on just the density and the density gradient.

In the context of generalized gradient approximations, the exchange functional is usually expressed in terms of the electron density, ρ , and its gradient, $\nabla\rho$, as a simple product of the local spin density approximation (LSDA) and an enhancement factor:

$$E_x^{\text{GGA}} = \int d^3r \rho \varepsilon_x^{\text{LSDA}}(\rho) F_x^{\text{GGA}}(s) \quad (1)$$

where (all equations are in Hartree atomic units)

$$\varepsilon_x^{\text{LSDA}} = - (3/4)(3/\pi)^{1/3} \rho^{1/3} \quad (2)$$

is the exchange energy density per particle for a uniform electron gas (UEG), F_x^{GGA} is the enhancement factor, and

$$s = |\nabla\rho|/[2(3\pi^2)^{1/3} \rho^{4/3}] \quad (3)$$

is the variable called the dimensionless reduced gradient, defined on the interval $s \in [0, \infty)$.

All of the formulas are presented for a closed-shell system, where we can dispense with the spin components; the extension to open shells is standard and summarized in the Appendix. The approximation called LSDA here reduces to the local density approximation (LDA) for closed-shell Slater determinants with all orbitals doubly occupied, but we call it LSDA in all of our publications because that is the more general case, of which LDA

is a special case, and it would be confusing to use both names for this kind of functional (the description of all functionals becomes simpler for closed-shell systems).

As suggested by Becke,² it is convenient to perform a change of variable from the variable s to a new finite variable

$$u_\gamma = \frac{\gamma s^2}{1 + \gamma s^2}, u_\gamma \in [0, 1] \quad (4)$$

and to define the enhancement factor as a function of the new variable, where γ is a constant. We will use a special case u of this variable with $\gamma = 1$:

$$u = \frac{s^2}{1 + s^2} \quad (5)$$

and eq 1 can be expressed in the new variable as

$$E_x^{\text{GGA}} = \int d^3r \rho \varepsilon_x^{\text{LSDA}}(\rho) F_x^{\text{GGA}}(u) \quad (6)$$

Then, the exchange energy is completely determined by the shape of the curve $F(u)$ in the region $u \in [0, 1]$.

The shape of $F_x^{\text{GGA}}(s)$ or $F_x^{\text{GGA}}(u)$ can be expressed in many different ways, the more common ones being simple functions of s (as in the PBE,³ RPBE,⁴ and B88⁵ functionals), or polynomial expansions (as in the B97² and SOGGA11⁶ functional), with coefficients determined by some combination of physical constraints and fitting to experimental or theoretical data or both. In general the flexibility of $F(s)$ or equivalently $F(u)$ is limited by the chosen functional form. Similar but more complicated considerations apply to the correlation functional.

An alternate way to define a continuous and differentiable $F_x^{\text{GGA}}(u)$ curve is to fix some points (knots) and then use a spline to interpolate them. If a cubic spline is used, then by construction

Received: September 3, 2011

Published: October 27, 2011

it provides continuous first and second derivatives. Higher-order splines or splines under tension can also be used. An advantage of using splines to represent $F_x^{\text{GGA}}(u)$ is that the exchange-correlation functional is not constrained by preconceptions about the functional form.

The present article has two parts. In the first part, we show that the spline representation of density functionals is not just a theoretical possibility but that it actually works well in practice. In the second part, we use a spline implementation of exchange and correlation functionals to show the sensitivity of the accuracy of several properties to specific regions of u (and therefore also, by eq 5, to specific regions of s).

2. DATABASE

The database used in this article comprises 12 subsets (smaller databases) corresponding to various chemical properties. All subsets are the same as used in our recent SOGGA11 paper,⁶ where they were mainly based on previous work.^{7–25} The properties of the subsets are as follows: main-group atomization energies (MGAE109/05),¹² ionization potentials (IP13/03),^{10,12–14} electron affinities (EA13/03),^{10,12–14} proton affinities (PA8/06),²² alkyl bond dissociation energies (ABDE12),^{6,12,15,16} difficult hydrocarbon cases (HC7/11),^{6,23,24} barrier heights (HTBH38/08 and NHTBH38/08),^{12,17,18,25} noncovalent interactions (NCCE31/05),^{11,13,21,24} atomic energies (AE17),⁸ metal bond energies (SRMBE12 and MRMBES),^{6,19,20} and main-group bond lengths (MGBL19).^{9,15}

The database is composed of 303 Born–Oppenheimer energies (that is, electronic energies including nuclear repulsion) or relative energies and 19 bond lengths. For example, for bond energies, we use equilibrium dissociation energies (D_e), not ground-vibrational-state dissociation energies (D_0) or bond enthalpies at finite temperature; the equilibrium dissociation energy is the difference between the Born–Oppenheimer energy at the dissociation limit and the equilibrium geometry. Relative energy data (including dissociation energies, barrier heights, overall energies of reactions, ionization potentials, van der Waals binding energies, etc.) require more than one single-point energy calculation, and therefore the total number of single-point calculations performed for the energetic part of the database is 350.

Geometries are the same as used previously.⁶

Next, we give more details of the database subsets.

2.1. Main Group Atomization Energies (MGAE109/05). The MGAE109/05 database⁶ consists of 109 atomization energies (AEs) for main-group compounds. We always give the mean errors in atomization energies on a per bond basis because that makes comparison between different test sets more portable. To make it possible for readers to convert to the mean unsigned error per molecule, we always compute the mean errors in atomization energies by computing the mean error per molecule and then dividing by the average number of bonds per molecule in the test set; the latter value is 4.71 for MGAE109/05. Geometries for all molecules in this database are obtained with the QCISD/MG3 method.^{7,26,27}

2.2. Ionization Potentials, Electron Affinities, and Proton Affinities (IP13/03, EA13/03, PA8/06). The zero-point-exclusive adiabatic ionization potential (IP) and electron affinity (EA) test sets are called IP13/03 and EA13/03, respectively, and they have been explained and employed in our previous papers.^{10,12–14} Geometries for both neutral and ionic species in

the IP13/03 and EA13/03 databases are optimized separately by QCISD/MG3;^{9,21} i.e., these are adiabatic, not vertical, IPs and EAs. PA8/06²² is a database of eight zero-point-exclusive proton affinities. Geometries for the PA8/06 database are obtained with the MP2(full)/6-311G(2df,p) method.^{28,29}

2.3. Alkyl Bond Dissociation Energy (ABDE12). The two databases for alkyl bond dissociation energies, ABDE4/05 and ABDEL8, were found to behave similarly in preliminary work, and we joined them in ABDE12. The ABDE4/05 database^{12,15,16} contains four bond dissociation energies of small R–X organic molecules, with R = methyl and isopropyl and X = CH₃ and OCH₃. ABDEL8⁶ contains a set of eight R–X bond dissociation energies including larger molecules, with R = ethyl and *tert*-butyl and X = H, CH₃, OCH₃, OH. For the 12 considered bonds, D_0 values are taken from a paper by Izgorodina et al.¹⁶ and converted to reference D_e values by using B3LYP/6-31G(d) zero-point vibrational energies scaled with a scale factor of 0.9806.

2.4. Hydrocarbons Difficult Cases (HC7/11). The HC7 database^{23,24} consists of seven difficult cases involving medium-range correlation energies in hydrocarbons. HC7 is the combination of the HC5 database with two isodesmic reactions (involving adamantane and bicyclo[2.2.2]octane) that were singled out as difficult cases by Grimme. All geometries were obtained with the MP2/6-311+G(d,p) method.^{28,29} The original reference data for this database have been published in a previous paper,²⁴ and some inconsistencies were recently corrected.⁶

2.5. Barrier Heights (HTBH38/08, NHTBH38/08). The HTBH38/08 database contains 38 transition state barrier heights for 19 hydrogen transfer (HT) reactions, 18 of which involve radicals as reactants and products.^{12,17,18} NHTBH38/08 consists of three databases containing 38 transition state barrier heights for non-hydrogen-transfer (NHT) reactions. The individual databases contain 12 barrier heights for heavy-atom transfer reactions, 16 barrier heights for nucleophilic substitution (NS) reactions, and 10 barrier heights for non-NS unimolecular and association reactions. Fourteen data points in this database were revised in 2008²⁵ and collected with other data in the DBH24/08 subset. Any of the data that were updated in DBH24/08 were updated in these databases, which are now renamed HTBH38/08 and NHTBH38/08. All geometries are obtained with the QCISD/MG3 method.^{7,26,27}

2.6. Noncovalent Interaction (NCCE31/05). Several databases have been developed in our group for various kinds of noncovalent interactions, in particular, HB6/04,²¹ CT7/04,²¹ DI6/04,²¹ WI7/05,¹³ and PPS5/05.¹³ HB6/04 is a hydrogen bond database that consists of the equilibrium binding energies of six hydrogen-bonded dimers. The CT7/04 database consists of binding energies of seven charge transfer complexes. The DI6/04 database contains the binding energies of six dipole interaction complexes. The WI7/05 database consists of the binding energies of seven weak interaction complexes, all of which are bound by dispersion-like interactions. The PPS5/05 database consists of binding energies of five π – π stacking complexes. We used them here grouped in the NCCE31/05 database.²⁴ The geometries for the benzene dimers in the NCCE31/05 database are taken from Sinnokrot and Sherrill,¹¹ while geometries for all other molecules in this database are optimized with the MC-QCISD/3 method.³⁰

2.7. Atomic Energies (AE17). AE17 is composed of 17 total atomic energies of the atoms from H to Cl.⁸

2.8. Metals (SRMBE12, MRMBE5). In a recent paper,⁶ we rearranged the databases^{19,20} related to metal bond energies by dividing them into single-reference metal bond energies and multireference metal bond energies. We keep this new division in this paper, and we use the SRMBE12 and MRMBE5 subsets as presented in ref 6.

2.9. Main Group Bond Lengths (MGBL19). MGBL19 is a database of 19 experimental bond lengths of 15 small main-group molecules,¹⁵ and the experimental bond lengths are taken from a previous compilation by Hamprecht et al.⁹

3. COMPUTATIONAL DETAILS

The MG3S basis set¹⁰ and the ultrafine (99,590) Lebedev grid are used throughout the paper for all calculations. Restricted calculations were performed for closed-shell species, while unrestricted calculations were employed for open-shell species. In systems like the oxygen atom, each orbital is an eigenfunction of the single-electron orbital angular momentum operators L^2 and L_z . However, singly occupied p orbitals have different radial functions and orbital energies than doubly occupied ones, and the atomic Slater determinants are not eigenfunctions of the many-electron orbital angular momentum operators L^2 and L_z . All calculations in section 4 are exchange-only (no correlation), but all those in section 5 include correlation.

4. THE DFT-SPLINE IMPLEMENTATION

4.1. Exchange. The key component in the DFT-spline implementation is the change of variables from s to u , given by eq 5, and the inverse transformation given by

$$s = \sqrt{\frac{u}{1-u}} \quad (7)$$

We replace the canonical enhancement factor of eq 6 with a natural cubic spline, f_x :

$$E_x^{\text{GGA}} = \int d^3r \rho \varepsilon_x^{\text{LSDA}}(\rho) f_x(u) \quad (8)$$

Given that a sufficient number of knots are provided to allow the spline to accurately describe the original curve, this implementation is capable of replicating the results of any GGA functional that can be expressed as eq 6.

We define our spline as a piecewise cubic polynomial that interpolates a series of j points at equidistant knots, u_j . A total number of n_{spline} knots are taken, with separation $u = 1/(n_{\text{spline}} - 1)$ in the interval $u \in [0, 1]$. The spline is defined as³¹

$$f_x = A_j f_j + B_j f_{j+1} + C_j f_j'' + D_j f_{j+1}'', u_j \leq u \leq u_{j+1} \quad (9)$$

with

$$A_j = \frac{u_{j+1} - u}{n_{\text{spline}}} \quad (10)$$

$$B_j = 1 - A = \frac{u - u_j}{n_{\text{spline}}} \quad (11)$$

$$C_j = \frac{1}{6}(A_j^3 - A_j)(1/n_{\text{spline}})^2 \quad (12)$$

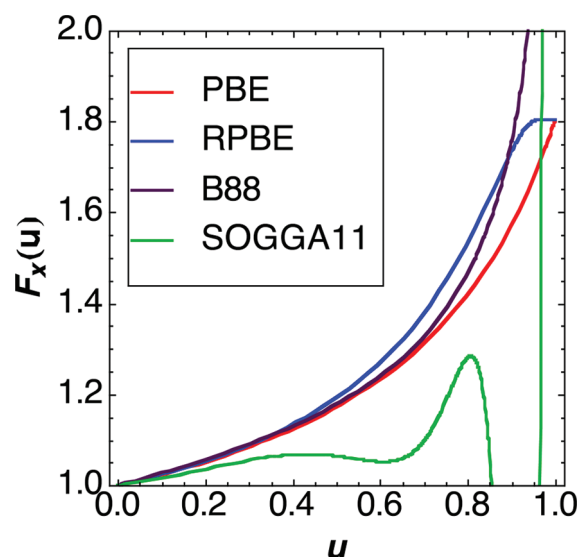


Figure 1. Enhancement factors for different exchange functionals in the new variable u .

$$D_j = \frac{1}{6}(B_j^3 - B_j)(1/n_{\text{spline}})^2 \quad (13)$$

$$u_j = (j - 1)/(n_{\text{spline}} - 1) \quad (14)$$

and natural conditions on second derivatives at the end points

$$f_1'' = f_{n_{\text{spline}}}'' = 0 \quad (15)$$

The coefficients of the spline are calculated once, by means of a tridiagonal matrix algorithm based on Gaussian elimination³¹ and then stored. The value of the spline and its first derivative are then calculated on the fly from the stored coefficients at each point of the DFT integration grid. This results in an efficient algorithm, which is not more expensive than evaluating conventional exchange functionals. The evaluation of the enhancement factor in a standard DFT implementation is a negligible part of the cost in a DFT calculation, and the spline implementation retains this advantage.

4.2. Implementation of Common GGA Functionals by Means of the DFT-Spline Algorithm. In this work, we implemented some of the most successful GGA exchange functionals by using the DFT-spline algorithm in a modified version of the Gaussian 09 program.³² The functionals considered are PBE,³ RPBE,⁴ B88,⁵ and SOGGA11.⁶

The four enhancement factors are shown in Figure 1, and their formulations are obtained from the original definitions of $F(s)$ by applying the transformation in eq 7. The enhancement factors in the transformed variable were used to calculate the ordinates at equidistant knots of the abscissa u . The DFT-spline algorithm was used to interpolate the points, and we expect its performances to be directly influenced by the number of the knots used by the algorithm.

As pointed out in the original paper, SOGGA11 has an oscillating behavior for large u . Although the spline implementation allows an easy fix for such a problem, this is not the purpose of the present work.

We studied the convergence of the results obtained with the DFT-spline implementation as a function of the total number of

Table 1. Root Mean Square Deviation (RMSD/ E_h)^a from the Absolute Atomic Energies Obtained with the Usual Implementation as a Function of the Number of the Knots in the Spline (n_{spline}) for Four Exchange Functionals

n_{spline}	PBE (XO) ^b	RPBE (XO) ^b	B88 (XO) ^b	SOGGA11 (XO) ^b
11	1×10^{-5}	2×10^{-4}	1×10^{-3}	2×10^{-2}
21	9×10^{-7}	8×10^{-6}	5×10^{-4}	6×10^{-3}
101	2×10^{-9}	8×10^{-9}	5×10^{-6}	5×10^{-6}
201	3×10^{-10}	4×10^{-10}	2×10^{-7}	6×10^{-8}
1001	$\leq 1 \times 10^{-10}$	$\leq 1 \times 10^{-10}$	1×10^{-8}	2×10^{-9}
2001	$\leq 1 \times 10^{-10}$	$\leq 1 \times 10^{-10}$	$\leq 1 \times 10^{-10}$	$\leq 1 \times 10^{-10}$

^a $1 E_h = 1$ hartree = 27.2114 eV. ^b XO denotes that these are exchange-only calculations (no correlation)

knots, n_{spline} , for these four exchange functionals by comparing them to those obtained from the usual implementation of the same functionals. For this study, we used exchange-only density functionals to calculate SCF atomic energies of the first 17 atoms (H to Cl), performing unrestricted calculations with the MG3S²² basis set. The root-mean-square deviations (RMSDs) in hartrees from the energies obtained with the usual implementation are presented as a function of n_{spline} in Table 1.

The results in Table 1 show that the spline implementation is capable of replicating conventional results for all considered functionals. In general, there is good agreement between the atomic energies obtained with the DFT-spline implementation and the usual implementation. For PBE and RPBE, RMSDs smaller than 2×10^{-4} are obtained with the surprisingly small value of $n_{\text{spline}} = 11$. For more complicated functionals, like B88 and the SOGGA11 exchange, a slightly larger number of points is necessary to reach similar convergence. At the extreme value of $n_{\text{spline}} = 2001$, the atomic energies calculated with the DFT-spline and the canonical implementation are identical within $10^{-10} E_h$. However, we find that $n_{\text{spline}} = 101$ (corresponding to one point every 0.01 u units) provides a RMSD smaller than $5 \times 10^{-5} E_h$ for all functionals, and this value is sufficient for the description of atomic energies to chemical accuracy.

In order to confirm these results on molecular energies, we used the large database described in section 2. For each of the 350 atomic and molecular energy calculations required to compute the energetic part of the database, we calculated the difference between the absolute energies obtained with the DFT-spline implementation and those obtained from the usual implementation without fitting. These calculations are performed by combining the DFT-spline implementation of the exchange with $n_{\text{spline}} = 101$ with correlation functionals calculated using the usual implementation. The B88 exchange functional was coupled with the LYP correlation to obtain BLYP energies. PBE and RPBE were coupled with PBE correlation, and SOGGA11 exchange was coupled with SOGGA11 correlation. The RMSDs are presented in Table 2, and they confirm the excellent performance of the DFT-spline implementation in the calculations of the considered functionals. In particular, all of the results on the larger molecular database are very similar to those obtained for the smaller atomic set using the same number of points, $n_{\text{spline}} = 101$.

As a further validation of the performance of the new implementation, we used the DFT-spline results to calculate the mean unsigned errors of the 303 energetic chemical data in the considered databases. Results for the MUE of these data do

Table 2. Root Mean Square Deviation (RMSD/ E_h) of Results Obtained Using $n_{\text{spline}} = 101$ from the Absolute Energies Obtained Without Fitting for the 350 Atomic and Molecular Single-Point Energies in the Database^a

functional	RMSD
PBE	5×10^{-9}
RPBE	1×10^{-8}
BLYP	5×10^{-6}
SOGGA11	7×10^{-6}

^a These calculations include both exchange and correlation, but the spline is applied only to exchange in this table.

not differ from those obtained with the nonfitted implementation up to a level of precision of 10^{-3} kcal/mol.

5. SENSITIVITY STUDY

The dependence of the exchange-correlation functional on s or u plays a crucial role in the development of density functionals since it entirely determines the performance of a GGA functional. Values at some knots are fixed by physical constraints, e.g., the uniform electron gas fixes the value of the gradient enhancement factor to be 1 at $u = 0$. The known second order coefficient for the density gradient expansion fixes the value of the first derivative with respect to u at $u = 0$, but the global u dependence is unknown. It is generally believed that the significant region for the description of chemical properties lies between $0 \leq s \leq 3$,^{33,34} corresponding to u in the range 0–0.9, but more detailed information about which s or u region contributes most to each property is hard to obtain. An advantage of the DFT-spline implementation lies in the fact that it provides a simple tool to investigate the shape of the enhancement factor as a function of the reduced density gradient variable.

A natural cubic spline implementation for correlation is not as straightforward as the one for exchange because the correlation enhancement factors generally depend on a spin variable ζ as well as the density and density gradient (see the Appendix). In order to perform the sensitivity study, we modify the exchange or correlation component of the GGA functionals under study by multiplying by a spline factor, in particular

$$E_x^{\text{GGA}} = \int d^3r \varepsilon_x^{\text{GGA}}(\rho, s) f_x(u) \quad (16)$$

for exchange and

$$E_c^{\text{GGA}} = \int d^3r \varepsilon_c^{\text{GGA}}(\rho, s, \zeta) f_c(u) \quad (17)$$

for correlation, where f_x and f_c are both natural cubic splines of the form of eq 9. The base spline is a trivial one obtained by setting the value to $f = 1$ for both the exchange and the correlation at all 21 knots. The sensitivity analysis to different ranges is performed by modifying the base spline by either increasing or decreasing the value of f at one single knot, $j + 1$, by 5%, in order to obtain a perturbation at the corresponding value of u .

Examples of the resulting splines are shown in Figure 2, where the first and ninth points are brought from 1.0 to 1.05 to obtain f_+^j and from 1.0 to 0.95 to obtain f_-^j . The 5% value was chosen because it is large enough for the perturbation to be significant but is still small enough to avoid extensive ringing near the perturbed point (as can be seen in Figure 2). For each modification

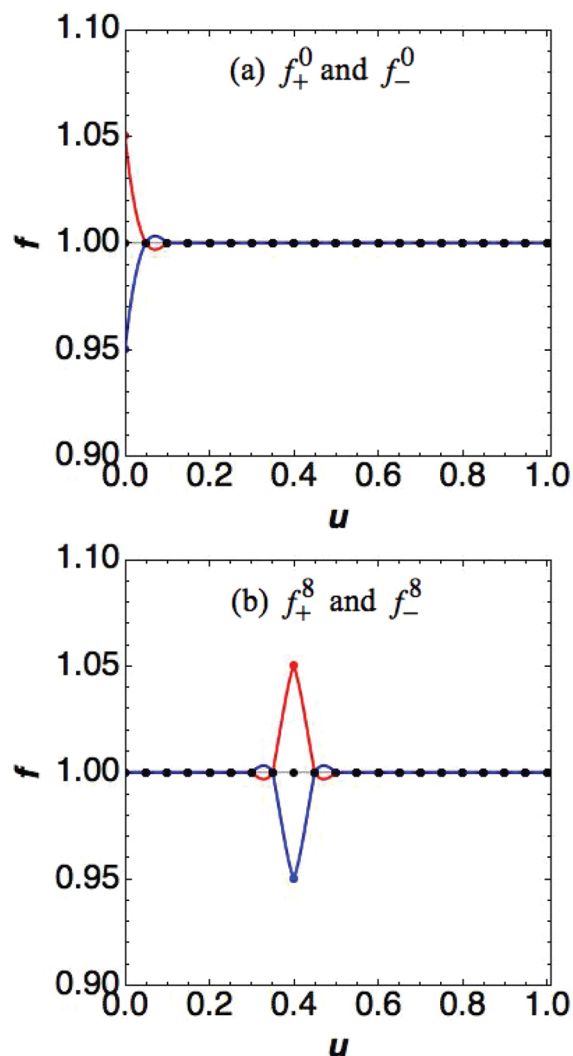


Figure 2. Perturbation of the base spline (black) to form (a) f_+^0 (red) and f_-^0 (blue) and (b) f_+^8 (red) and f_-^8 (blue).

of each exchange or correlation functional, we performed full SCF calculations on the chosen database.

For every modification of the base spline, we calculate the sensitivity (S) of the results by using the following formula:

$$S = \frac{\Delta^+ + \Delta^-}{2} \quad (18)$$

with

$$\Delta^\pm = \pm \frac{1}{\delta} \frac{\text{MUE}(\pm \delta) - \text{MUE}(0)}{\text{MUE}(0)} \quad (19)$$

where δ is the relative perturbation applied (± 0.05 in our case). Then Δ^\pm measures the sensitivity of the properties as a function of the perturbation applied, for example, if a positive perturbation of 5% in u decreases the MUE by 10%, we have $\Delta^+ = -0.10/0.05 = -2$. At the same time, if a negative perturbation in u worsens the results by 5%, we have $\Delta^- = 0.05/-0.05 = -1$. The sensitivity in this case will be $S = (-2 - 1)/2 = -1.5$.

Results for the sensitivity study of the four considered exchange functionals and two correlation functionals are discussed in the next sections. PBE and SOGGA11 were chosen for

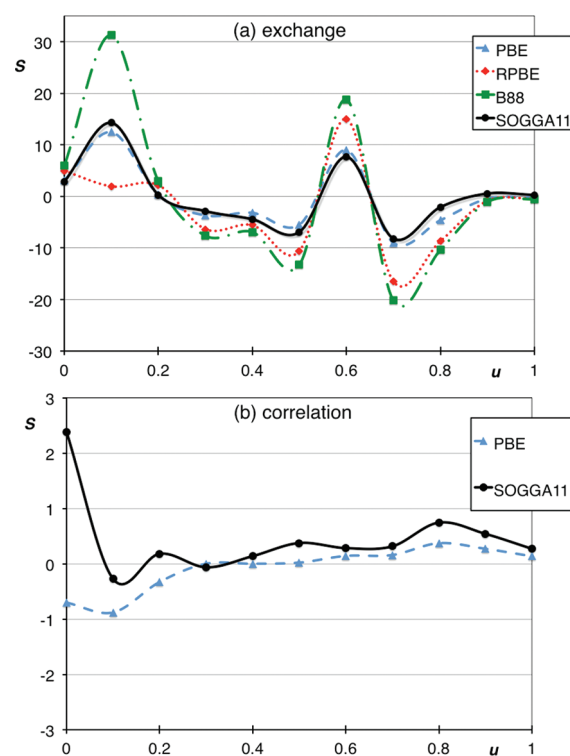


Figure 3. Sensitivity study to variation of (a) f_x and (b) f_c as a function of u for MGAE109/05.

study as representative correlation functionals. All four exchange functionals and both correlation functionals studied here satisfy the uniform electron gas constraint at $u = 0$.

The correlation energy is on average an order of magnitude smaller than the exchange energy, and effects of a given percentage change in the correlation functional therefore tend to be much smaller than those for changing the exchange functional. For these reasons, all plots for correlation are reported with a smaller scale.

5.1. Main Group Atomization Energies (MGAE109/05).

The effects of the perturbations at localized regions of u on the main-group atomization database are collected in the parts a and b of Figure 3. All four of the exchange functionals are sensitive to changes over almost all of the values of u . We can distinguish two regions of importance, one around $u = 0.1$ and another between $u = 0.5$ and $u = 0.8$. It is interesting to note a significant sensitivity at $u = 0.8$ ($s = 2$) and even higher, which corresponds to a region that was previously thought to have less significant importance.

For the correlation functional, the results are mainly sensitive in the small u region, including the UEG point at $u = 0$. For correlation too, we find an unexpected sensitivity at large values of u , around $u = 0.8$.

In order to make these results more concrete and less abstract, let us consider an example. Consider the PBE functional; the MUE per bond for PBE with this database is 2.99 kcal/mol. Decreasing the enhancement factor by 5% at $u = 0.1$ changes this to 1.69 kcal/mol (a 44% decrease), whereas increasing it by 5% at $u = 0.1$ changes the MUE to 5.41 kcal/mol (an 81% increase). Therefore, S is the average of 16.2 and 8.7, which yields the plotted value of 12.5. In general, a positive S means that the accuracy is increased by decreasing the density functional at that point.

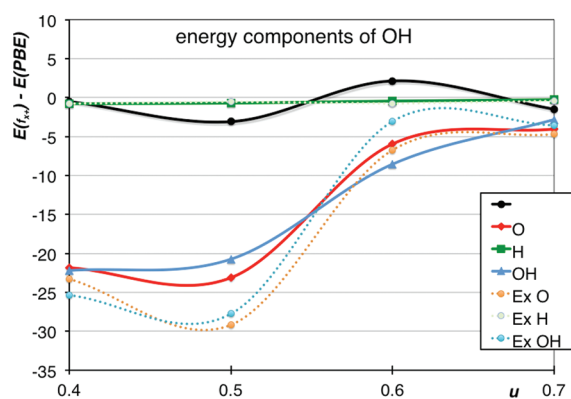


Figure 4. Analysis of the changes in the energy components of the OH radical and its bond dissociation products upon increasing f_x from 1.0 to 1.05 at selected values of u . The curves labeled Ex show changes in exchange energy of O, H, and OH. The curves labeled O, H, and OH show changes in the total energy of the atoms and molecule. The black curve is the change in bond energy (total energy of the atoms minus total energy of the molecule).

To illustrate the origin of the different behavior as a function of u , we pick the OH radical as a representative molecule from the MGAE109/05 database. The PBE functional has an error of 2.4 (to make it easier to read, all energies in the rest of this discussion are expressed in kcal/mol and rounded to the nearest tenth) for the OH bond energy (accurate, 107.1; unaltered PBE, 109.5), which is similar to the mean unsigned error of 3.0, and its behavior around $u = 0.6$ is also similar to the average over the whole database.

For the OH molecule, the oxygen atom, and the hydrogen atom, we plot in Figure 4 the differences between the absolute energies (solid) and the exchange energies (dotted) obtained using the base spline and that obtained using the perturbed spline f_{x+}^j with 5% positive perturbations in the range $0.4 \leq u \leq 0.7$. The plot also shows the change in the bond energy.

The differences for the hydrogen atom are not significant for both the exchange (changes from -0.4 to -0.8 kcal/mol) and the total energies (changes from -0.3 to -0.8), and it is clear that the changing in behavior of the atomization energy curve is given mainly by the different sensitivity of the OH molecule and the O atom. Furthermore, although the correlation energy does change (because the SCF calculation with a perturbed exchange functional leads to a perturbed density), in all cases (atoms and molecules) the change is small (usually of magnitude less than 0.1, always of magnitude less than 0.2). This shows that the change in the direct energy contribution of the density functional is mainly due to the change in exchange energy of O and OH.

The exchange energy of O before the perturbation is -5112.7 , while that of OH is -5347.7 . The contribution of exchange to the bond energy (exchange energy of O plus H minus that of OH) is 44.9 kcal/mol.

Since the perturbation is positive, the exchange energy increases in magnitude, becoming more negative. For a perturbation at $u = 0.6$, the exchange energy of O becomes -5119.5 , a change of -6.7 , while that of OH becomes -5350.7 , a change of only -3.1 . The contribution of exchange to the bond energy becomes 40.5 kcal/mol, a perturbation of the exchange contribution to the bond energy of -4.4 . However, the perturbation changes the orbitals. While the total energy of O changes

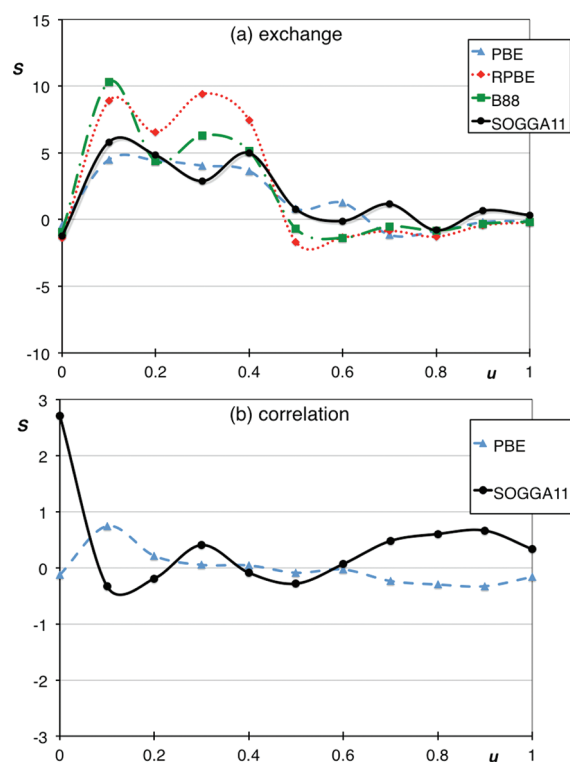


Figure 5. Sensitivity study to variation of (a) f_x and (b) f_c as a function of u for IP13/03.

by -6.0 , that of OH is more sensitive, and it changes by -8.6 . The net change to the bond energy is $+2.2$, raising the error from 2.2 to 4.6. A negative perturbation produces changes of a similar magnitude but in the opposite direction. The large magnitude of the individual terms illustrates the well-known fact that quantum mechanics involves large cancellations of absolute energy contributions to produce chemical results like bond energies, and the competing effects of this example make it more understandable that the sign of S can change as a function of u . For example, as seen in the figure, the exchange curves of O and OH cross so that a positive perturbation at $u = 0.4$ changes the exchange energy of O by -23.3 and that of OH by -25.3 . Hence, the exchange contribution to the bond energy is perturbed in the opposite direction of that at $u = 0.6$, and the bond energy change is also in the opposite direction. The purpose of the sensitivity analysis is not to trace all of these details through the SCF calculations and the changes in orbitals but rather to see if changes in the exchange and correlation functional in certain regions of u have systematic effects on the accuracy of prototype chemical properties, so that one knows which regions of u space are important for more systematic improvement of exchange–correlation functionals for better chemical predictions in the future.

5.2. Ionization Potentials (IP13/03). Results for the IP13/03 database are collected in Figure 5. Ionization potentials are sensitive to the small u region of the exchange functional, $0.1 \leq u \leq 0.4$. The errors are most sensitive to changes in the correlation functional at $u = 0$ and at $u \geq 0.7$.

5.3. Electron Affinities (EA13/03). Results for the EA13/03 database are collected in Figure 6. The effects of changes in the exchange for electron affinities are similar to those for ionization

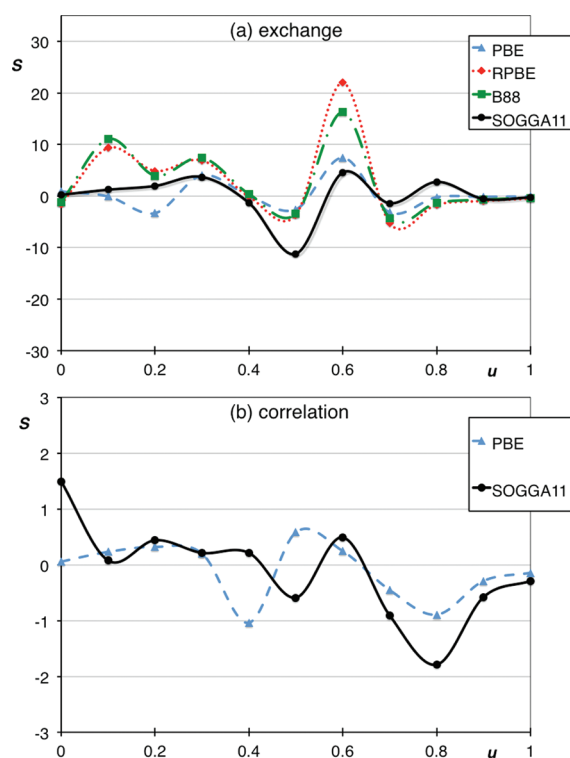


Figure 6. Effects of the variation of (a) f_x and (b) f_c as a function of u on the MUE of EA13/03.

potentials, with a broad sensitive region between $0.1 \leq u \leq 0.3$. A key difference in these two properties is the region around $u = 0.7$, which is more important for the electron affinities than for the ionization potentials.

For the correlation, we notice sensitive effects in regions where the exchange is less affected, in particular $u = 0.8$.

Comparing Figures 3–6, an interesting conclusion emerges. In particular, all four exchange functionals could all be improved for all three properties by decreasing the exchange enhancement factor around $u = 0.6$. However, decreasing the SOGGA correlation function at $u = 0.8$ improves the atomization energies and IPs but makes the EAs worse. The latter shows the limitations of GGAs; they are not flexible enough to fit all properties, but the former shows the inflexibility of prechosen functional forms, not of GGAs per se.

5.4. Proton Affinities (PA8/06). Results for proton affinities are presented in Figure 7. This property has less sensitivity than ionization potentials and electron affinities, with the most significant centered at about $u = 0.1$ for both the exchange and the correlation.

5.5. Alkyl Bond Dissociation Energies (ABDE12). Results for alkyl bond dissociation energies are reported in Figure 8.

Results for the exchange are affected in two main regions, one at $u = 0.1$ and the other one at $u = 0.5–0.7$. It is interesting to note that all functionals seem improvable at $u = 0.1$, but if the goal is to obtain a good overall functional (as it was for SOGGA11), performances in that region for ABDE12 must be balanced with those for MGAE109/05, which is also very sensitive here but in the opposite direction. This shows why it is important to have both atomization energies and bond energies in training and test sets.

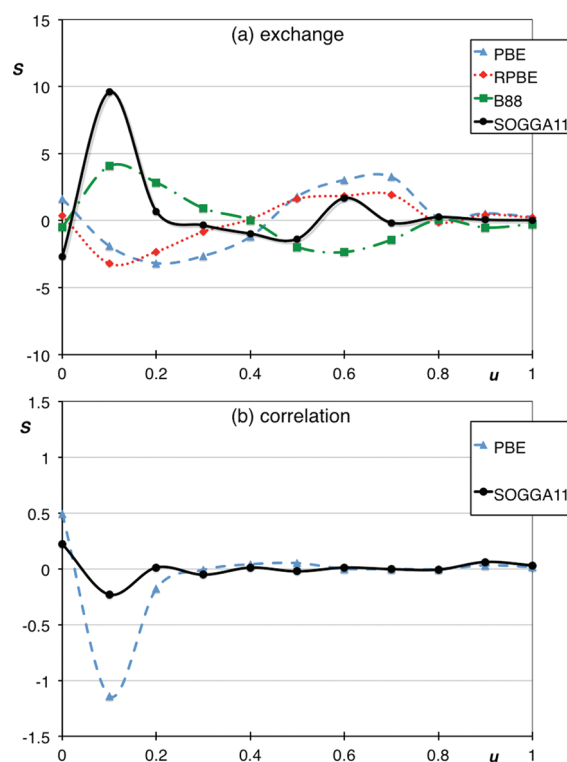


Figure 7. Effects of the variation of (a) f_x and (b) f_c as a function of u on the MUE of PA8/06.

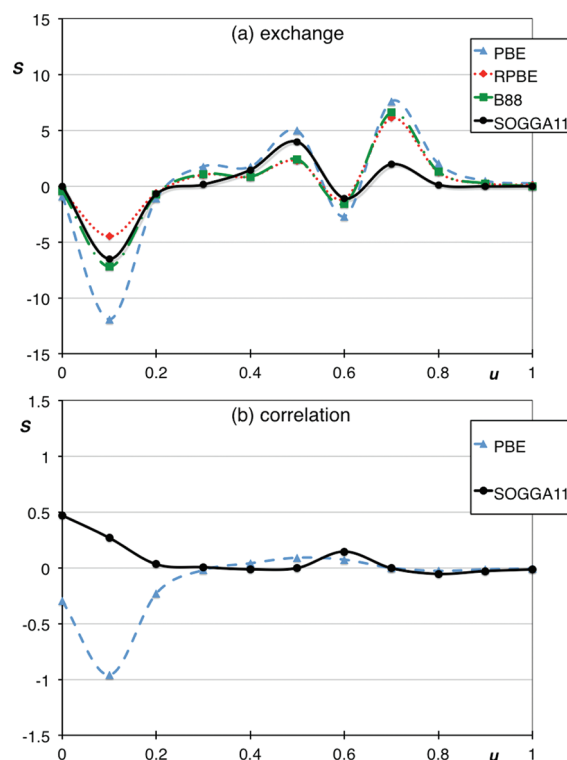


Figure 8. Effects of the variation of (a) f_x and (b) f_c as a function of u on the MUE of ABDE12.

Correlation effects are once again smaller than those for exchange, with a sensitive region for alkyl bond energies at $u = 0.1$.

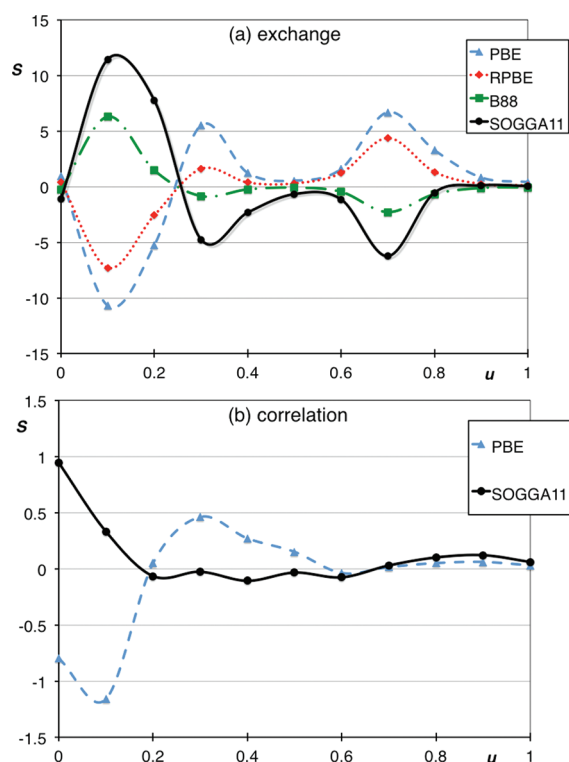


Figure 9. Effects of the variation of (a) f_x and (b) f_c as a function of u on the MUE of HC7.

5.6. Difficult Hydrocarbon Cases (HC7/11). Results for the difficult hydrocarbon cases are presented in Figure 9.

The most sensitive regions for the exchange are again the regions around $u = 0.1$ and around $u = 0.7$. This implies that improving GGA functionals for this property must be carefully balanced with the performances for atomization energies and alkyl bond dissociation energies.

The sensitivity to changes in the correlation functional is very small for $u \leq 0.6$ and almost zero at larger values of u .

5.7. Barrier Heights (HTBH38/08, NHTBH38/08). Results for the HTBH38/08 database are collected in Figure 10, and results for NHTBH38/08 are presented in Figure 11. Barrier heights have long been a difficult problem for DFT, and these figures help us to understand why.

Consider first the more important exchange functional. Two of the exchange functionals have positive sensitivity at $u = 0.1$ for both kinds of barriers, but the other two show opposite sensitivity. For the two functionals with opposite signs of S , improving the performance for one kind of barrier will worsen it for the other. Furthermore, for all four functionals, the sensitivity fluctuates wildly (as a function of u) for NHTBH38/08. Restricted analytic forms may not have the flexibility to be optimum for all regions of u .

Barrier heights are much less sensitive to the correlation functional, confirming what we expected from previous experience. The present results are more definitive though, since they do not suffer from the possible inadequacy of choosing an insufficiently flexible functional form.

5.8. Noncovalent Interactions (NCCE31/05). Sensitivity results for noncovalent interactions are presented in Figure 12. The sensitivity of this property to the exchange functional is spread out over a wide range of u . Interestingly, the

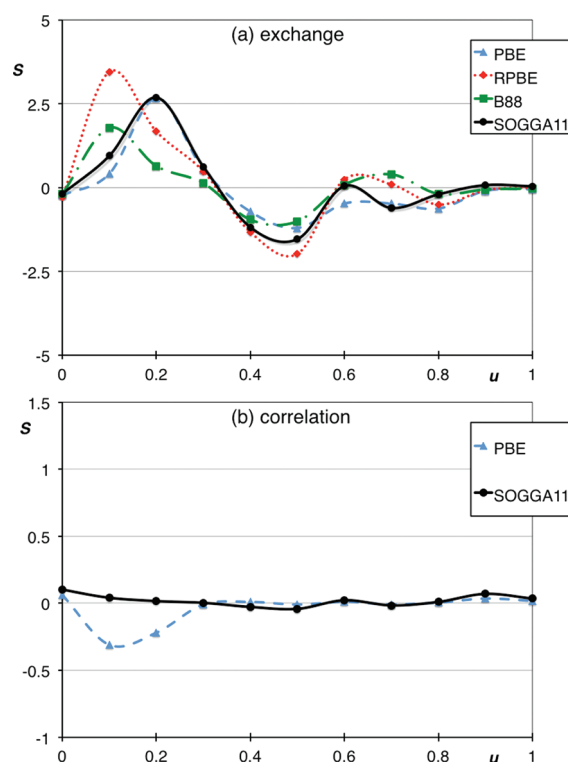


Figure 10. Effects of the variation of (a) f_x and (b) f_c as a function of u on the MUE of HTBH38/08.

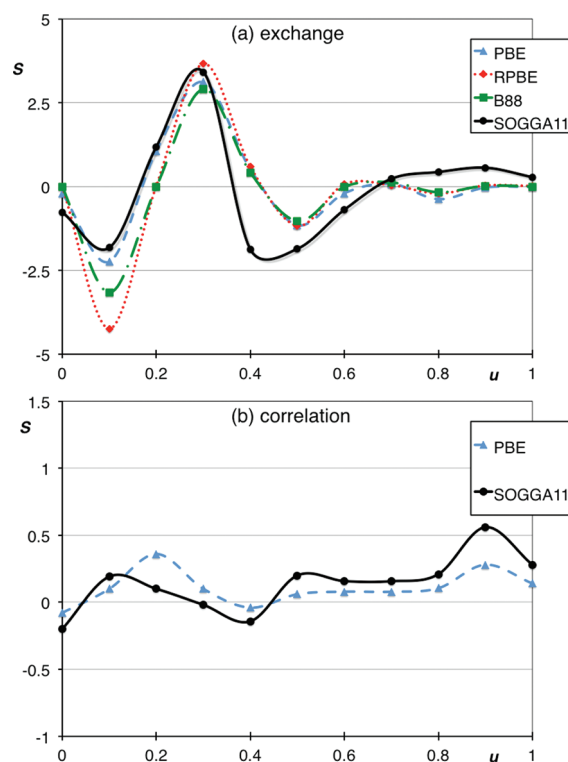


Figure 11. Effects of the variation of (a) f_x and (b) f_c as a function of u on the MUE of NHTBH38/08.

sensitivity to the exchange functional is smallest for SOGGA11, perhaps indicating that this functional is close to optimal

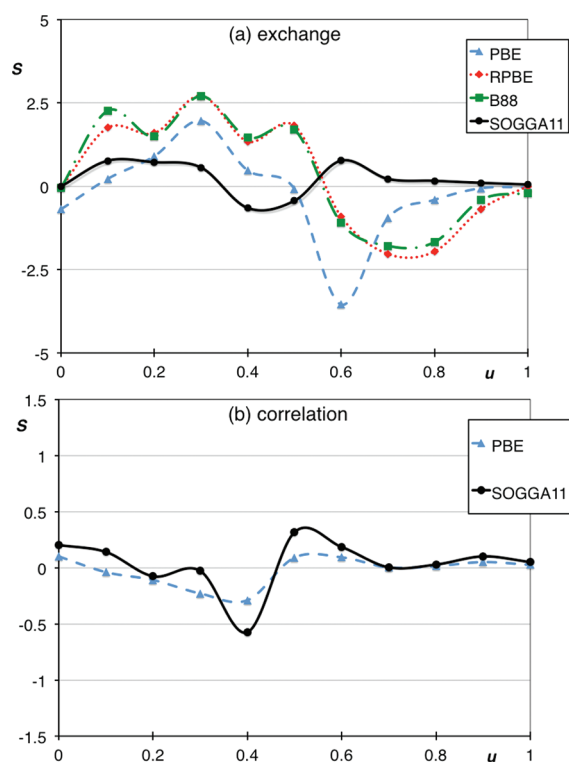


Figure 12. Effects of the variation of (a) f_x and (b) f_c as functions of u on the MUE of NCCE31/05.

(the first derivative with respect to changes vanishes for an optimum function).

From a more general perspective, it is known that the behavior of the enhancement factor at large values of the reduced density gradient s affects the performance of the exchange functional for noncovalent interactions.^{35–38} For example, functionals (such as B88) with enhancement factors that approach large values or infinity at large s have a very different performance for noncovalent interactions than functionals (such as PBE) with enhancement factors that asymptotically reach a small value. This fact is clearly reflected in the mean signed errors of each group of functionals for the noncovalent interaction data set. For example, B88 underestimates the attractive noncovalent interactions by about 3 kcal/mol, while PBE overestimates them by about 1 kcal/mol. Our sensitivity analysis, however, tests the performance of each functional for small perturbations of the enhancement factor, and our plots, like all sensitivity analyses based on small perturbations, are incapable of showing this behavior. This is one of the disadvantages of sensitivity analysis in general.

In a recent paper, Johnson et al.,³⁸ by analysis of the electron density, found by another route that noncovalent interactions are indeed sensitive to a large range of s . Their analysis included larger molecules of biological interest, with similar conclusions to those drawn from our sensitivity analysis, which, however, adds the important observation that most of the sensitivity is coming from the exchange functional. Correlation contributes to the sensitivity in the region around $u = 0.4–0.5$, corresponding to $s = 0.8–1$, but the effect is smaller than for exchange.

One encouraging result is that for the exchange functional at $u = 0.3$. Figures 3–11 show that the sensitivities of the PBE, RPBE, and B88 functionals are all positive or small at this u value

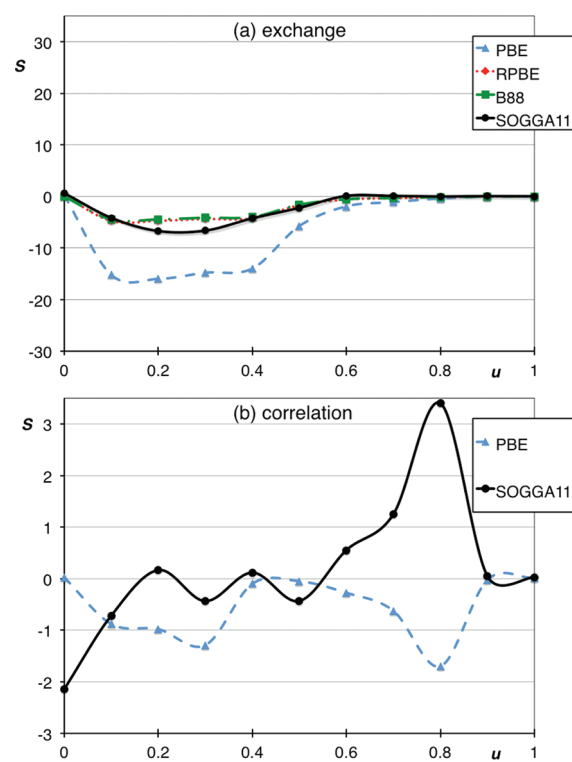


Figure 13. Effects of the variation of (a) f_x and (b) f_c as functions of u on the MUE of AE17.

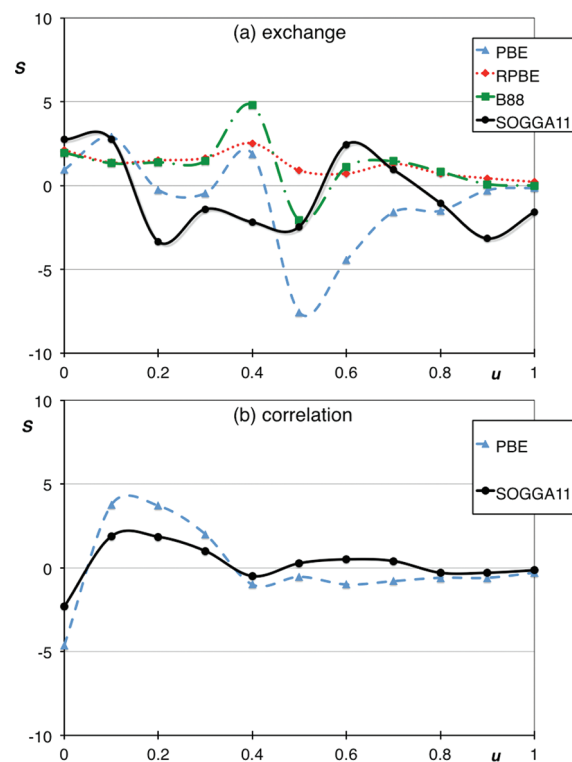


Figure 14. Effects of the variation of (a) f_x and (b) f_c as a function of u on the MUE of SRMBE12.

for the wide range of nonmetallic molecular properties represented in these nine figures. Thus, decreasing the exchange at this

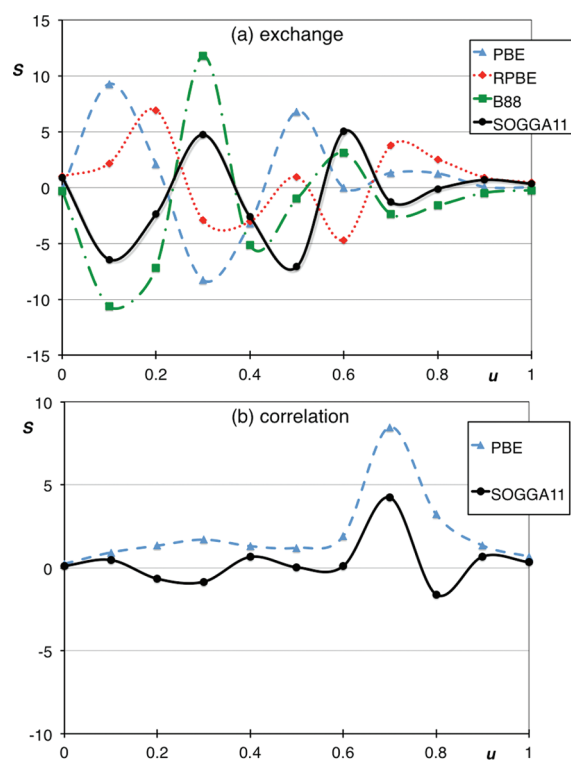


Figure 15. Effects of the variation of (a) f_x and (b) f_c as a function of u on the MUE of MRMBE5.

u value should improve those functionals broadly. If we had not already optimized the exchange and found that these functionals have too much exchange at this u value (see Figure 1), this might have led us to discover this. However, for SOGGA11, S is negative in Figure 9, indicating that some values would improve and some would get worse if it were decreased further.

5.9. Atomic Energies (AE17). Results for the 17 total atomic energies of the atoms from H to Cl are collected in Figure 13. Atomic energies show a broad sensitive region for the PBE functional in the region $0.1 \leq u \leq 0.5$; this is an indication of systematic error in that the functional could be improved by increasing the exchange. However, as is clear from the discussion in the previous subsection, increasing the exchange functional in that region would make many molecular properties worse, at the expense of improving atomic absolute energies.

At $u = 0$ and at $u = 0.8$, there is basically no sensitivity for all exchange functionals, while the correlation functional is most sensitive in these regions.

5.10. Metals (SRMBE12, MRMBE5). The results for single-reference metal bond energies (SRMBE12) and the multi-reference metal bond energies (MRMBE5) are reported in Figures 14 and 15.

Compounds containing transition metals are difficult to treat with DFT, and the results for the exchange confirm this prediction. It is hard to find common ground from the sensitivity study, but in general we can say that SRMBE12 is less sensitive to changes in the exchange functional than is MRMBE5, whereas the opposite is true for correlation. SRMBE12 shows sensitivity in the entire range of u . Both correlation functionals show sensitivity in the small u region for SRMBE12 and a pronounced sensitivity around $u = 0.7$ for MRMBE5.

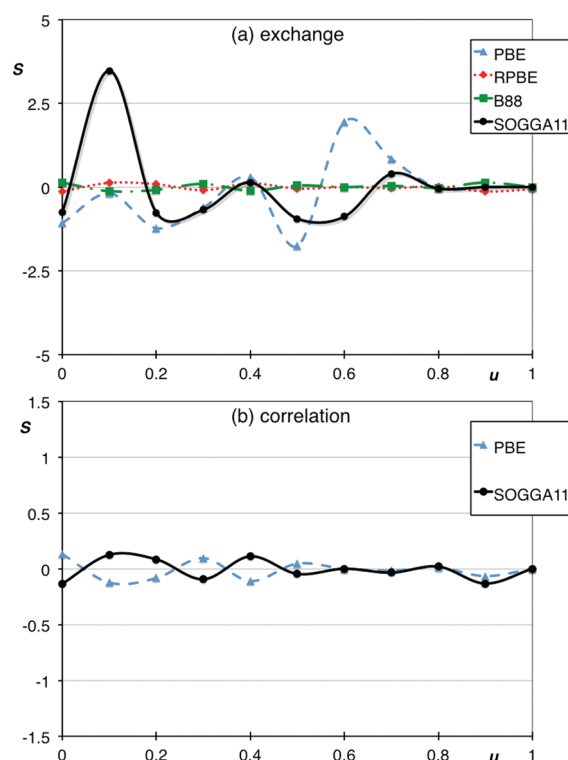


Figure 16. Effects of the variation of (a) f_x and (b) f_c as a function of u on the MUE of MGBL19.

The different behavior of the SOGGA11 sensitivity in Figure 14 as compared to the other three can be understood to some extent by the behaviors shown in Figure 1, except at $u = 0.6$. However, Figure 15 is very hard to understand.

5.11. Main Group Bond Lengths (MGBL19). Results on the geometries database MGBL19 are reported in Figure 16. The sensitivity of the bond length to changes in the exchange are quite small, with two interesting regions at $u = 0.1$ and $u = 0.6$. The effects of the correlation are also very small.

6. CONCLUDING REMARKS

We presented in this article an implementation of four common GGA exchange functionals by means of a natural cubic spline. We showed that the spline is able to reproduce the unfitted results very accurately. This opens the door to future work in which density functionals can be defined directly in terms of splines, without being constrained by guessed functional forms. Other advantages of the DFT-spline implementation include very convenient calculation of the first and second derivatives, offering a simpler implementation in quantum chemistry codes for any GGA functional, and the possibility to create new fitted functionals that are not subject to Runge's phenomenon of oscillation for high-order interpolating polynomial functions.

By using multiplicative factors defined in terms of spline functions, we were able to study the exchange and correlation components of GGAs to various regions of the reduced density for 12 energetic chemical databases and one bond-length database. This task was hardly possible before, with the usual DFT implementation. In general, if we consider the results for the

exchange, we usually find reasonably good agreement for all considered functionals as to which are the sensitive regions of u . The two correlation functionals are, however, usually sensitive to other regions.

In some cases, the sensitivity analysis shows the limitations of GGAs in that changes of the density functionals in a given direction make one property better and another worse. This is particularly clear for the exchange functional in the range $0 \leq u \leq 0.5$. For example, when a negative perturbation is applied in that range, properties such as MGAE109/05, HC7, PA8/06, NCCE31/05, and MGBL19 often improve, presenting a positive sensitivity, but properties such as ABDEL12 and AE17 often worsen, with negative sensitivity. Barrier heights are less sensitive than other properties, but hydrogen transfers (HTBH38/08) and non-hydrogen transfers (NHTBH38/08) present different sensitivities to changing the exchange in the same region. At large values of u , sensitivity is always reduced compared to the small u region, but some databases such as IP13/03, PA8/06, HTBH38/08, NHTBH38/08, and AE17 present basically no sensitivity to variations in the exchange at large u , while the others present some sensitivity, with contrasting behaviors.

Perturbation of the correlation functional always produces smaller effects, but also in this case we can find examples of different behaviors in the $0 \leq u \leq 0.5$ range, for example, positive values of S for properties such as MGAE109/05, IP13/03, and SRMBE12 and negative values for EA13/03 and PA8/06. Thus, any attempt at optimization involves a trade-off where some quantities get better and others get worse. We find that some databases, e.g., MGAE109/05, PA8/06, ABDEL12, HC7, HTBH38/08, NCCE31/05, SRMBE12, and MGBL19, do not have significant sensitivity at large values of u , while others have small but non-negligible sensitivity. In particular, it is interesting to note the different behavior of the two barrier height databases (NHTBH38/08 and HTBH38/08) and the large sensitivity of AE17 in this large u region, as opposed to what was found for exchange.

APPENDIX: FORMULAS FOR THE SPIN-UNRESTRICTED CASE

Exchange. The exchange energy for a spin-polarized system ($\rho \uparrow \neq \rho \downarrow$) is evaluated in the usual way from the exchange functional for a spin-unpolarized system ($\rho \uparrow = \rho \downarrow$) by using the spin-scaling relation:

$$E_x\{\rho \uparrow, \rho \downarrow\} = E_x[2\rho \uparrow]/2 + E_x[2\rho \downarrow]/2 \quad (\text{A.1})$$

where $E_x[\rho] \equiv E_x\{\rho/2, \rho/2\}$ is calculated with eq 1 from the main text.

Correlation. The spin-polarized case for correlation is more complicated, and various approaches are used in the literature. The PBE and SOGGA11 correlation functionals (which are the only ones considered here) use an enhancement factor that depends upon the relative spin polarization $\zeta = (\rho \uparrow - \rho \downarrow)/\rho$:

$$E_c\{\rho \uparrow, \rho \downarrow\} = \int d^3r \rho \{\epsilon_c^{\text{LSDA}}(\rho, \zeta) + H(\rho, u, \zeta)\} \quad (\text{A.2})$$

AUTHOR INFORMATION

Corresponding Author

*E-mail: truhlar@umn.edu.

ACKNOWLEDGMENT

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

REFERENCES

- (1) Kohn, W.; Becke, A.; Parr, Y. J. *Phys. Chem.* **1996**, *100*, 12974.
- (2) Becke, A. D. *J. Chem. Phys.* **1996**, *107*, 8554–8560.
- (3) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1995**, *77*, 3865–3868.
- (4) Hammer, B.; Hansen, L.; Norskov, J. *Phys. Rev. B* **1999**, *59*, 7413–7421.
- (5) Becke, A. D. *Phys. Rev. A* **1987**, *38*, 3098–3100.
- (6) Peverati, R.; Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. Lett.* **2011**, 1991–1997.
- (7) Pople, J.; Head-Gordon, M.; Raghavachari, K. *J. Chem. Phys.* **1987**, *87*, 5968–5975.
- (8) Chakravorty, S.; Gwaltney, S.; Davidson, E.; Parpia, F.; Fischer, C. *Phys. Rev. A* **1993**, *47*, 3649–3670.
- (9) Hamprecht, F.; Cohen, A.; Tozer, D. J.; Handy, N. C. *J. Chem. Phys.* **1997**, *109*, 6264–6271.
- (10) Lynch, B. J.; Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2003**, *107*, 1384–1388.
- (11) Sinnokrot, M. O.; Sherrill, C. D. *J. Phys. Chem. A* **2003**, *108*, 10200–10207.
- (12) Zhao, Y.; Schultz, N. E.; Truhlar, D. G. *J. Chem. Theory Comput.* **2005**, *2*, 364–382.
- (13) Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2005**, *109*, 5656–5667.
- (14) Zhao, Y.; Schultz, N. E.; Truhlar, D. G. *J. Chem. Phys.* **2005**, *123*, 161103.
- (15) Zhao, Y.; Truhlar, D. G. *J. Chem. Phys.* **2005**, *125*, 194101.
- (16) Izgorodina, E.; Coote, M.; Radom, L. *J. Phys. Chem. A* **2005**, *109*, 7558–7566.
- (17) Zhao, Y.; Lynch, B. J.; Truhlar, D. G. *Phys. Chem. Chem. Phys.* **2005**, *7*, 43–52.
- (18) Zhao, Y.; González-García, N.; Truhlar, D. G. *J. Phys. Chem. A* **2005**, *109*, 2012–2018.
- (19) Schultz, N. E.; Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2005**, *109*, 4388–4403.
- (20) Schultz, N. E.; Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2005**, *109*, 11127–11143.
- (21) Zhao, Y.; Truhlar, D. G. *J. Chem. Theory Comput.* **2005**, *1*, 415–432.
- (22) Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2006**, *110*, 10478–10486.
- (23) Zhao, Y.; Truhlar, D. G. *Org. Lett.* **2006**, *8*, 5753–5755.
- (24) Zhao, Y.; Truhlar, D. G. *Theor. Chem. Acc.* **2008**, *120*, 215–241.
- (25) Zheng, J.; Zhao, Y.; Truhlar, D. G. *J. Chem. Theory Comput.* **2009**, *5*, 808–821.
- (26) Curtiss, L.; Raghavachari, K.; Redfern, P.; Rassolov, V.; Pople, J. *J. Chem. Phys.* **1997**, *109*, 7764–7776.
- (27) Fast, P.; Sanchez, M.; Truhlar, D. *Chem. Phys. Lett.* **1999**, *306*, 407–410.
- (28) Moller, C.; Plesset, M. *Phys. Rev.* **1933**, *46*, 0618–0622.
- (29) Radom, L.; Schleyer, P. R.; Pople, J. A.; Hehre, W. J. *Ab Initio Molecular Orbital Theory*, 1st ed.; Wiley: New York, 1986.
- (30) Lynch, B.; Truhlar, D. *J. Phys. Chem. A* **2003**, *107*, 3898–3906.
- (31) de Boor, C. *A Practical Guide to Splines*, Rev. ed.; Springer-Verlag: New York, 2001.
- (32) 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.; Ogario, 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.1; Gaussian, Inc.: Wallingford, CT, 2009.

(33) Zupan, A.; Perdew, J.; Burke, K. *Int. J. Quantum Chem.* **1997**, *61*, 835–845.

(34) Zupan, A.; Burke, K.; Emzerhof, M.; Perdew, J. *J. Chem. Phys.* **1997**, *106*, 10184–10193.

(35) Lacks, D.; Gordon, R. *Phys. Rev. A* **1993**, *47*, 4681–4690.

(36) Zhang, Y.; Pan, W.; Yang, W. *J. Chem. Phys.* **1997**, *107*, 7921.

(37) Kannemann, F. O.; Becke, A. D. *J. Chem. Theory Comput.* **2009**, *5*, 719–727.

(38) Johnson, E. R.; Keinan, S.; Mori-Sanchez, P.; Contreras-Garcia, J.; Cohen, A. J.; Yang, W. *J. Am. Chem. Soc.* **2010**, *132*, 6498–6506.

Assembling Small Silicon Clusters Using Criteria of Maximum Matching of the Fukui Functions

Edison Osorio,[†] Marta B. Ferraro,[‡] Ofelia B. Oña,[‡] Carlos Cardenas,[§] Patricio Fuentealba,[§] and William Tiznado^{*,†}

[†]Departamento de Química, Facultad de Ciencias Exactas, Universidad Andres Bello, Avenida República 252, Santiago de Chile

[‡]Departamento de Física, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, and IFIBA, CONICET, (1428) Ciudad Universitaria - Pab. I., Buenos Aires, Argentina

[§]Departamento de Física, Facultad de Ciencias, Universidad de Chile, Casilla 653, Santiago, Chile and CEDENA

ABSTRACT: In this work, we present a methodology inspired by criteria of “maximum matching” between the Fukui functions to predict the best interaction between small silicon clusters to form larger ones. The model is based on the topological analysis of the Fukui functions. We tested the methodology in the formation of Si₄–Si₈ using a set of small Si₂–Si₆ clusters as building blocks in ground state structures in singlet and triplet multiplicities. In all of the cases, the Fukui function predicts the formation of the large cluster in its ground state structure, but the number of reaction channels increases with the cluster size.

INTRODUCTION

Since the advent of advanced laser vaporization techniques,^{1–5} atomic and nanocluster semiconductors have become an active subject of research, both experimentally and theoretically.^{6–10} Spectroscopic experiments have been utilized to infer the most stable structures of small atomic clusters, but measurements provide only indirect information about the geometries. Moreover, there is not conclusive experimental evidence on what is the mechanism of formation of the more stable structures in atomic clusters. Silicon nanoclusters have been extensively studied because of their intrinsic interest from the point of view of chemical structure and bonding as well as their importance in the microelectronics industry.^{11,12} Motivated by the fundamental importance of understanding properties of silicon-based materials with increasing size, especially the transition, atom → cluster → bulk, considerable experimental and theoretical efforts have been devoted to determine geometric structures of small and midsize silicon clusters.^{8,13–15}

In the past three decades the development of chemical-reactivity density functional theory^{16–20} has provided a formal framework for many empirical chemical concepts like electronegativity,²¹ hardness,^{22–26} Fukui function,^{27–32} electrophilicity,^{33–35} et cetera. These descriptors have been widely applied to study chemical reactivity in organic chemistry, less in inorganic chemistry, and little in solid state surfaces^{36,37} and clusters.^{38–40}

The aim of this paper is to use a criterion of “maximum matching” between the Fukui functions of small clusters to predict how large clusters can be formed from the small ones. This criterion indicates that where electrostatic effects are not decisive, at the offset of the cluster formation, the clusters prefer to orient such that their Fukui functions maximize their overlap. To check this criterion, we assemble pairs of small silicon clusters in a “maximum matching” orientation and then relax the structure and check whether the final geometry corresponds to a known stable structure of the resulting cluster. The organization of the present paper is as follows. In the next section, we will

address some formal issues like the approximation used to calculate the Fukui function and the information obtained from their topological analysis and how the attractors of the Fukui functions are used to predict the most efficient interaction between two small clusters to produce a larger one, the “maximum matching” criteria. The Fukui function information is used to predict the most favorable interaction of the small clusters in each considered reaction; the efficiency of the Fukui function in predicting the most stable isomers is discussed.

THEORETICAL MODEL

In the density-functional theory approach to phenomenological chemical reactivity theory (chemical DFT),^{16,20,41,42} the reactive site of an acceptor of electrons is associated with a large positive value of the Fukui function:^{29,30}

$$f^+(\mathbf{r}) = \left(\frac{\partial \rho(\mathbf{r})}{\partial N} \right)_{v(\mathbf{r})}^+ = \rho_{N+1}(\mathbf{r}) - \rho_N(\mathbf{r}) \quad (1)$$

The superscript “+” on the derivative indicates that the derivative is taken from above; this is essential because the derivative from above and the derivative from below

$$f^-(\mathbf{r}) = \left(\frac{\partial \rho(\mathbf{r})}{\partial N} \right)_{v(\mathbf{r})}^- = \rho_N(\mathbf{r}) - \rho_{N-1}(\mathbf{r}) \quad (2)$$

are not equal when the number of electrons is an integer, owing to the derivative discontinuity of the energy,^{43,44} density, and other molecular properties.^{31,45} Similarly, the Fukui function from below is the key regioselectivity indicator for the donor of electrons.

The link between the chemical DFT description and the frontier MO theory description is clear when one approximates

Received: September 14, 2011

Published: November 14, 2011

the Fukui functions using the frontier molecular orbitals:^{28,46}

$$\begin{aligned} f^-(\mathbf{r}) &= |\varphi_{\text{HOMO}}(\mathbf{r})|^2 \\ f^+(\mathbf{r}) &= |\varphi_{\text{LUMO}}(\mathbf{r})|^2 \end{aligned} \quad (3)$$

These approximations are sufficient except for the seemingly rare cases where orbital relaxation effects are important.^{47–49} When the HOMO or LUMO belongs to a degenerate irreducible representation of the cluster's point group, an average over the set of degenerated orbitals is used.^{50–52}

The following approximation for the total interaction energy of a donor and acceptor has been deduced elsewhere:^{25,53}

$$\begin{aligned} \Delta W_{\text{int}} &= (\mu_{\text{acceptor}} - \mu_{\text{donor}})\Delta N \\ &+ \int \left(\sum_{\alpha \in \text{donor}} Z_{\alpha} \delta(\mathbf{r} - \mathbf{R}_{\alpha}) - \rho_{\text{donor}}(\mathbf{r}) \right) \Phi_{\text{acceptor}}(\mathbf{r}) \, d\mathbf{r} \\ &+ \Delta N \int (f_{\text{acceptor}}^+(\mathbf{r}') \Phi_{\text{donor}}(\mathbf{r}) - f_{\text{donor}}^-(\mathbf{r}) \Phi_{\text{acceptor}}(\mathbf{r})) \, d\mathbf{r} \\ &- (\Delta N)^2 \iint \frac{f_{\text{donor}}^-(\mathbf{r}) f_{\text{acceptor}}^+(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} \, d\mathbf{r} \, d\mathbf{r}' \end{aligned} \quad (4)$$

where donor and acceptor stand for properties of the donor and acceptor, and the charge and position of the nuclei are indexed by α . The chemical potential, μ , determines who is the acceptor (the one with the most negative μ) and who is the donor (the one with less negative μ). ΔN is the number of electrons transferred from the donor to the acceptor. $\delta\nu_{\text{donor}}(\mathbf{r})$ is the change in the external potential of the acceptor due to the presence of the donor. $\delta(\mathbf{r} - \mathbf{R}_{\alpha})$ is the Dirac δ function centered at the position of the nucleus α , which accounts for the electrostatic interaction between the nuclei of both species, which are considered point charges. $\Phi(\mathbf{r})$ is the molecular electrostatic potential. The first term in eq 4 is independent of the orientation of the molecule and does not play a role in regioselectivity. The second term is a pure electrostatic interaction, and it is generally either small for reagents with noncharged sites (atoms) or negative for reagents with charged sites (atoms). In our case of neutral homonuclear silicon clusters, this term must be very small because the charge on Si atoms is negligible. The third term is the correction to the electrostatic interaction because of the electron transfer between both molecules. This term is usually positive: electron transfer weakens the strength of the electrostatic interaction because electron flow from the donor to the acceptor tends to equalize the charges of the reactive sites. This term, again, should be small in our case, as there is not significant charge to equalize. Finally, the integrand in the last term is usually positive because the Fukui function is positive, with some exceptions only in small regions around the nuclei of molecules with nodes of the frontier orbitals in the nuclear positions.^{49,54} This term, then, is stabilizing and depends entirely on the relative orientation between the “frontier” densities of the molecules;⁵⁵ this term is decisive for explaining the regioselectivity of systems like homonuclear clusters, where electrostatic interactions are expected to be negligible. Therefore, information on the optimal orientation between two small clusters to form a large one is enclosed in this term: the orientation should be such that the integrand is a maximum. This is what we called the *maximum matching* criterion. Summarizing, the relative

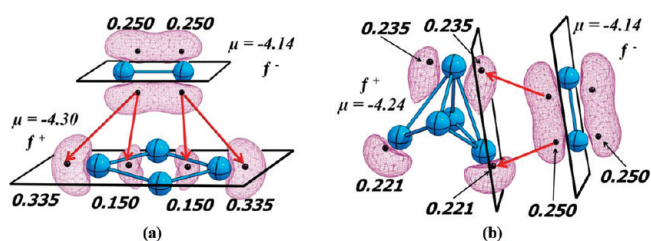


Figure 1. Interaction of Si_2^+ + Si_m^+ ($m = 4, 5$) guided by the respective Fukui functions. Black spheres correspond to attractors, and numbers correspond to the value of the Fukui function integrated in each basin.

orientation between reactants is optimal when

$$\iint \frac{f_{\text{donor}}^-(\mathbf{r}) f_{\text{acceptor}}^+(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} \, d\mathbf{r} \, d\mathbf{r}' \quad (5)$$

is maximum.

Optimizing eq 5 is not an easy task, as it not only depends on the orientation between the reactants but also on the distance between them. A systematic approach to this problem would be to fix a distance between the center of mass of both clusters and optimize the orientation between the molecules and then repeat for different distances until it is certain that the maximum had been achieved. However, for small clusters, one can approach the problem in an empirical way that uses information of the topology of the Fukui function gradient field.⁵⁶ Associated with the Fukui function, there is a gradient vector field, $\nabla f(\mathbf{r})$. This field is characterized by so-called critical points, where $\nabla f(\mathbf{r}) = (0, 0, 0)$. They represent local maxima, minima, and saddle points of $f(\mathbf{r})$. From a Hessian matrix, $H(f(\mathbf{r}_c))$, analysis associated with each critical point defines four different types of nondegenerate critical points: attractor (3, -3), repeller (3, 3), and saddle points (3, 1) and (3, -1). A basin, roughly speaking, is a region of the space, Ω , given by all points whose gradient paths end at the same attractor. The integral of the Fukui function in each basin, f_k , is a measure of the “abundance (population)” of the Fukui function around the attractor k . It is, then, reasonable to assume that at a given distance between the reactants (fragments), an assembling of the fragments that makes small the total distance between the attractors corresponding to the more populated basins (f_k large) translates into a large overlap of the Fukui functions.

The details and steps of the method proposed above are as follows:

- (i) The first step is to determine which cluster is expected to be the donor and which the acceptor. The acceptor between two clusters is the one with the smallest chemical potential (more negative).
- (ii) From a topological analysis of $\nabla f(\mathbf{r})$, all basins should be identified. Then, proceed with the integration of the Fukui function to obtain the set of f_k^\pm .
- (iii) In each cluster should be plotted planes which include as many attractors as possible with large f_k^\pm .
- (iv) The plane of the acceptor should be placed parallel to the plane of the donor “minimizing” the distance between attractors. In a symmetric cluster where the plane is the same as the molecular plane or a face in 3D clusters, the distance between planes is set to 1.0 Å; otherwise, it is set to 0.5 Å. Figure 1 shows how this procedure is done for the formation of Si_6^+ from the pairs $\{\text{Si}_2^+ + \text{Si}_5^+\}$.

	Structure	μ	f^+	f^-
Si_2^s C_∞		-4.98		
Si_2^t C_∞		-4.14		
Si_3^s C_{2V}		-4.52		
Si_3^t D_{3h}		-4.25		
Si_4^s D_{2h}		-4.64		
Si_4^t D_{2h}		-4.31		
Si_5^s D_{3h}		-4.35		
Si_5^t C_{2V}		-4.24		
Si_6^s C_{2V}		-4.27		
Si_6^t D_{4h}		-4.11		

Figure 2. Structures, chemical potential μ (eV), and donor and acceptor Fukui functions with their corresponding condensed values for Si_n ($n = 2-8$) global minimum structures of singlet and triplet configurations.

(v) Finally, from this starting point, the structures should be relaxed using an *ab initio* calculation.

In the present work, we study the formation of closed shell clusters. We have used the Fukui function information to predict the best interaction between fragments, and then we relaxed the systems fixing the multiplicity. Obeying the principle of spin conservation, we have combined the following: singlet + singlet \rightarrow singlet and triplet + triplet \rightarrow singlet. If we would like to explore higher spin states, we can fix the multiplicity before relaxing the formed cluster, for example: singlet + triplet \rightarrow triplet, triplet + triplet \rightarrow quintuplet, and/or exciting the clusters before relaxing.

This methodology is in total agreement with the working equations applied to use the local hard and soft acids and bases principle, which relies on matching of the local softness of the atoms in the fragments. These working equations use the local softness which corresponds to the softness of an atom in a molecule to estimate the maximum matching between both

fragments; if we consider the local softness as obtained by the Sf_k expression, where S is the global softness and f_k is the condensed Fukui function on atom k , then, it is the f_k value which determines the local reactivity of the system.⁴¹ In the present methodology, we used a different strategy to locally condense the Fukui function using a topological analysis to localize regions where this function is maximum (attractors), obtaining local reactive regions around these maximum values (basins), which are not necessarily atomic regions.

COMPUTATIONAL DETAILS

The most stable structures for Si_n ($n = 2-8$) were taken from the literature^{8,13,14} and reoptimized at the B3LYP^{57,58}/6-311+G(d,p)⁵⁹ level. Total energies were recalculated using CCSD-(T). All of these calculations were done using the Gaussian 03 program.⁶⁰ For the calculation of Fukui functions, we used the HF wave function of the CCSD(T)⁶¹ calculation. The

Table 1. Reaction Energies Associated with the Formation of Si_n ($n = 4-8$) Clusters^a

cluster	entry	reaction	reaction energy	ΔE_{rx}
Si_4^{s}	1	$\text{Si}_2^{\text{s}} + \text{Si}_2^{\text{s}} \rightarrow \text{Si}_4^{\text{s}}$	-146.1	25.1
	2	$\text{Si}_2^{\text{t}} + \text{Si}_2^{\text{t}} \rightarrow \text{Si}_4^{\text{s}}$	-121.0	0.0
Si_5^{s}	1	$\text{Si}_3^{\text{s}} + \text{Si}_2^{\text{s}} \rightarrow \text{Si}_5^{\text{s}}$	-124.2	11.9
	2	$\text{Si}_3^{\text{t}} + \text{Si}_2^{\text{t}} \rightarrow \text{Si}_5^{\text{s}}$	-112.3	0.0
Si_6^{s}	1	$\text{Si}_4^{\text{t}} + \text{Si}_2^{\text{t}} \rightarrow \text{Si}_6^{\text{s}}$	-127.3	11.6
	2	$\text{Si}_4^{\text{s}} + \text{Si}_2^{\text{s}} \rightarrow \text{Si}_6^{\text{s}}$	-121.0	10.4
	3	$\text{Si}_3^{\text{t}} + \text{Si}_3^{\text{t}} \rightarrow \text{Si}_6^{\text{s}}$	-116.9	6.3
	4	$\text{Si}_3^{\text{s}} + \text{Si}_3^{\text{s}} \rightarrow \text{Si}_6^{\text{s}}$	-106.2	0.0
Si_7^{s}	1	$\text{Si}_5^{\text{t}} + \text{Si}_2^{\text{t}} \rightarrow \text{Si}_7^{\text{s}}$	-133.7	28.0
	2	$\text{Si}_5^{\text{s}} + \text{Si}_2^{\text{s}} \rightarrow \text{Si}_7^{\text{s}}$	-130.7	5.4
	3	$\text{Si}_4^{\text{t}} + \text{Si}_3^{\text{t}} \rightarrow \text{Si}_7^{\text{s}}$	-128.3	3.0
	4	$\text{Si}_4^{\text{s}} + \text{Si}_3^{\text{s}} \rightarrow \text{Si}_7^{\text{s}}$	-105.7	0.0
Si_8^{s}	1	$\text{Si}_4^{\text{t}} + \text{Si}_4^{\text{t}} \rightarrow \text{Si}_8^{\text{s}}$	-108.4	37.6
	2	$\text{Si}_5^{\text{t}} + \text{Si}_3^{\text{t}} \rightarrow \text{Si}_8^{\text{s}}$	-103.5	21.1
	3	$\text{Si}_6^{\text{t}} + \text{Si}_2^{\text{t}} \rightarrow \text{Si}_8^{\text{s}}$	-102.0	12.6
	4	$\text{Si}_6^{\text{s}} + \text{Si}_2^{\text{s}} \rightarrow \text{Si}_8^{\text{s}}$	-95.8	6.4
	5	$\text{Si}_5^{\text{s}} + \text{Si}_3^{\text{s}} \rightarrow \text{Si}_8^{\text{s}}$	-87.3	4.9
	6	$\text{Si}_4^{\text{s}} + \text{Si}_4^{\text{s}} \rightarrow \text{Si}_8^{\text{s}}$	-70.8	0.0

^a Values of energy are in kcal mol^{-1} . In all cases, the reactants and products are the global minimum structures at the corresponding multiplicity.

topological analysis of the Fukui function was done with the DGrid 4.4 set of programs.⁶²

RESULTS AND DISCUSSION

Figure 2 shows the geometry, chemical potential μ (eV), Fukui functions isosurfaces at 0.003 au, and the f_k^{\pm} for each one of the small clusters Si_n ($n = 2-6$) used in the formation reactions in their two lowest spin multiplicities (singlet and triplet). All structures in singlet configuration are more stable than those in triplet configuration, except for the case of Si_2 , whose global minimum configuration is triplet.^{8,13,14} The chemical potential of the closed shell configurations is more negative than its open shell counterpart in all cases.

Table 1 presents the proposed reaction channels of the type $\text{Si}_x^{\alpha} + \text{Si}_y^{\alpha} \rightarrow \text{Si}_n$ where α denotes singlet or triplet configurations and $x + y = n$. As it is known that for $n = 4-8$, a Si_n cluster in singlet configuration is the most stable; only this product channel has been studied. The number of reactive channels increases with the size of the cluster. The most probable reaction channel will be the one with the most stable reactants because all of the reactions are exothermic, and the probability to populate the isomers higher in energy is low. Therefore, in Table 1, the relative energies ΔE_{rx} are taken with respect to the most probable reaction channels. Hence, Si_6 will be most probably formed by the combination of two Si_3 's in their singlet state. In the second column of Table 1, each combination to obtain the large cluster is numbered

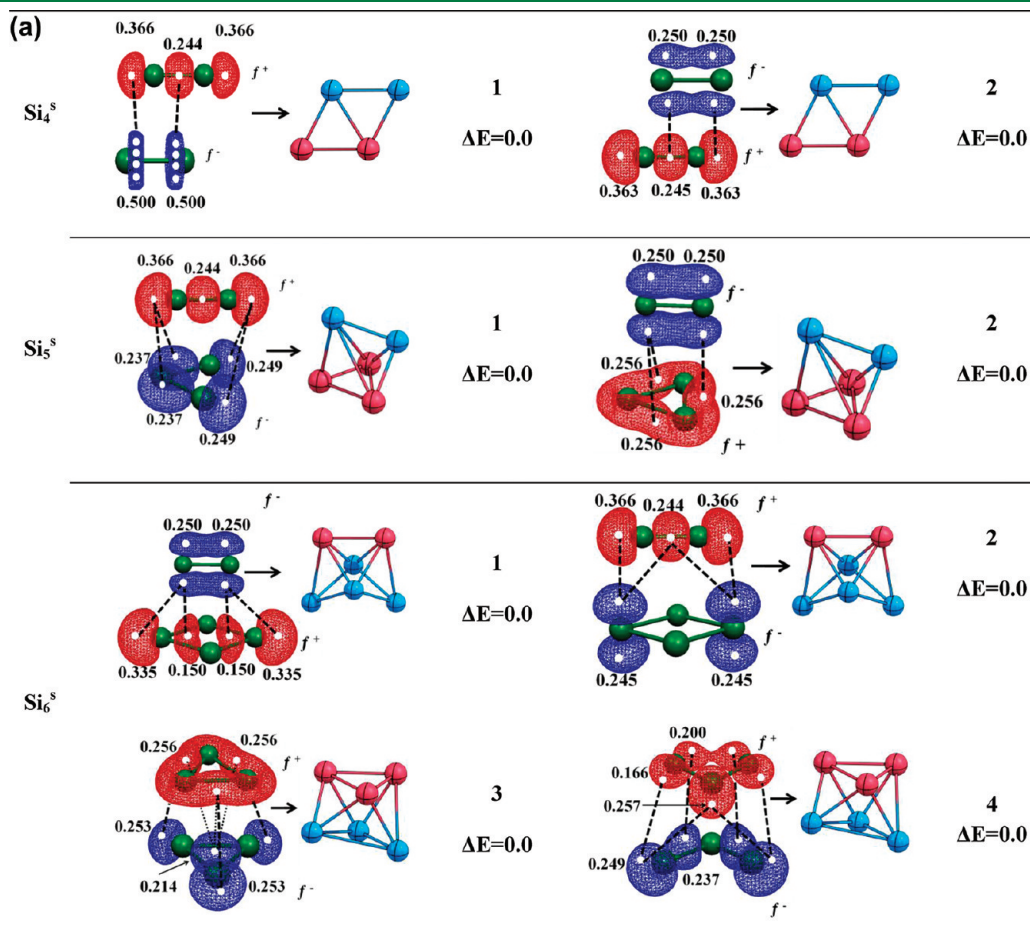


Figure 3. Continued

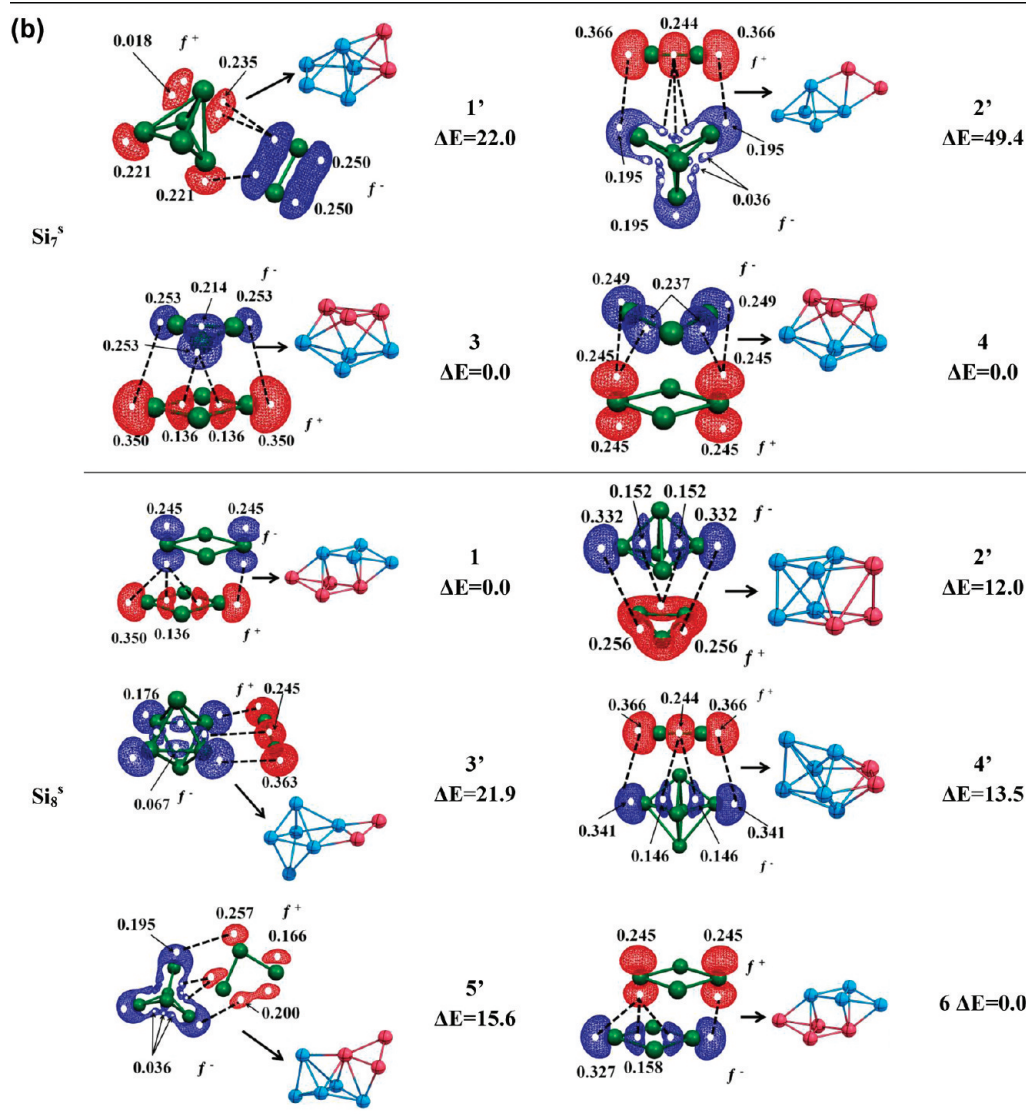


Figure 3. (a) Orientation that produces the maximum matching of the Fukui functions of two small clusters to form a large cluster (Si_4 – Si_6). f^+ (in red) is used for the clusters that accept electrons and f^- (in blue) for the donors. White dots represent the attractors' position of each Fukui function. The numbers represent the reaction channel according to Table 1, and ΔE is the isomer relative energy compared to the global minimum structure. (b) Orientation that produces the maximum matching of the Fukui functions of two small clusters to form a large cluster (Si_7 – Si_8). f^+ (in red) is used for the clusters that accept electrons and f^- (in blue) for the donors. White dots represent the attractors' position of each Fukui function. The numbers represent the reactions channel according to Table 1, and ΔE is the isomer relative energy compared to the global minimum structure. The prime identifies the cases where the global minimum structures are not obtained by the Fukui predictions.

(1, 2, ...) in a decreasing order of probability of the reaction to happen. It is important to remark that in Table 1, the reactions are proposed to produce the global minimum isomer.

Figure 3a and b show the position of the reactant clusters according to the maximum matching of the Fukui function. All of the predicted interactions lead to the formation of a stable cluster. The relative energies of the formation of clusters are also reported in Figure 3. All of the zero values correspond to cases where the most stable isomer was formed. In all of the cases the most stable reactants form the most stable product following the matching of the Fukui function criteria. There are cases where a cluster higher in energy is formed. For example, the Si_7 singlet formed following channels 1 and 2 of Table 1. It is also found that as the number of atoms in the cluster increases, a greater number

of low-lying energy isomers appear. This explains the variety of structures found in Si_8 . It is, therefore, gratifying to see that the criterion of maximum matching of the Fukui function is able to predict the most stable structure of Si_8 , which is the most challenging one among the studied cases.

CONCLUSIONS

The information obtained from topological analysis of the Fukui function, basins, attractors, and condensed values has been used to propose an empirical model to predict the best interaction between small silicon clusters to form larger clusters using a criterion of "maximum matching" between the Fukui functions. This criterion says that in reactions where electrostatic interactions are

negligible, two clusters prefer to orient such that the overlap, in the sense of eq 5, between their Fukui functions maximizes. Here, we show that in a significant number of examples of formation of small silicon clusters, this criterion successfully predicts the structure of the most stable isomers. In this study, we have not optimized the matching of the Fukui function in a rigorous way, but we have used information already contained in the gradient field of the Fukui functions. It is clear that in more complicated cases, as could be in systems with low symmetry or that are too big, an adequate maximization of eq 5 must be done. We have also devised how this can be done in a general way.

AUTHOR INFORMATION

Corresponding Author

*E-mail: wtiznado@unab.cl

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENT

Our work was supported by Andres Bello (Grant Nos. DI-13-10/R, DI-31-10/R, and DI-24-10/I) and Fondecyt (Grant Nos. 11090431, 1080184, 11090013). P.F. and C.C. also acknowledge the support by the Financiamiento Basal para Centros Científicos y Tecnológicos de Excelencia, Cedenna. M.B.F. and O.B.O acknowledge the support of Universidad de Buenos Aires (UBACYT 20020100100105) and CONICET (PIP0369).

REFERENCES

- (1) Kozlov, B.; Mamyrin, B. *Tech. Phys.* **1999**, *44*, 1073–1076.
- (2) Morales, A. M.; Lieber, C. M. *Science* **1998**, *279*, 208–211.
- (3) Oguri, K.; Okano, Y.; Nishikawa, T.; Nakano, H. *Phys. Rev. B* **2009**, *79*.
- (4) Popok, V. N.; Campbell, E. E. B. *Rev. Adv. Mater. Sci.* **2006**, *11*, 19–45.
- (5) Wegner, K.; Piseri, P.; Tafreshi, H. V.; Milani, P. *J. Phys. D: Appl. Phys.* **2006**, *39*, R439–R459.
- (6) Aiken, J. D.; Finke, R. G. *J. Mol. Catal. A: Chem.* **1999**, *145*, 1–44.
- (7) Bromann, K.; Félix, C.; Brune, H.; Harbich, W.; Monot, R.; Buttet, J.; Kern, K. *Science* **1996**, *274*, 956–958.
- (8) Nigam, S.; Majumder, C.; Kulshreshtha, S. K. *J. Chem. Phys.* **2006**, *125*.
- (9) Schwarz, W. H. E. *Angew. Chem.* **2006**, *118*, 1538–1547.
- (10) Sun, Z.; Yang, Z.; Gao, Z.; Tang, Z. C. *Rapid Commun. Mass Spectrom.* **2007**, *21*, 792–798.
- (11) Brown, W. L.; Freeman, R. R.; Raghavanchari, K.; Schlüter, M. *Science* **1987**, *235*, 860–865.
- (12) Schmid, G.; Hornyak, G. L. *Curr. Opin. Solid State Mater. Sci.* **1997**, *2*, 204–212.
- (13) Peppernick, S. J.; Gunaratne, K. D. D.; Sayres, S. G.; Castleman, A. W. *J. Chem. Phys.* **2010**, *132*.
- (14) Qin, W.; Lu, W. C.; Zhao, L. Z.; Zang, Q. J.; Wang, C. Z.; Ho, K. M. *J. Phys.: Condens. Matter* **2009**, *21*.
- (15) Rinnen, K. D.; Mandich, M. L. *Phys. Rev. Lett.* **1992**, *69*, 1823.
- (16) Ayers, P. W.; Anderson, J. S. M.; Bartolotti, L. J. *Int. J. Quantum Chem.* **2005**, *101*, 520–534.
- (17) Chermette, H. *Coord. Chem. Rev.* **1998**, *180*, 699–721.
- (18) Liu, S.-B. *Acta Phys.-Chim. Sin.* **2009**, *25*, 590–600.
- (19) Anderson, J. S. M.; Melin, J.; Ayers, P. W. *J. Chem. Theory Comput.* **2007**, *3*, 358–374.
- (20) Gazquez, J. J. *Mex. Chem. Soc.* **2008**, *52*, 3–10.
- (21) Parr, R. G.; Donnelly, R. A.; Levy, M.; Palke, W. E. *J. Chem. Phys.* **1978**, *68*, 3801–3807.
- (22) Parr, R. G.; Pearson, R. G. *J. Am. Chem. Soc.* **1983**, *105*, 7512–7516.
- (23) Pearson, R. G. *J. Am. Chem. Soc.* **1963**, *85*, 3533–3539.
- (24) Pearson, R. *J. Chem. Sci.* **2005**, *117*, 369–377.
- (25) Ayers, P. W. *Faraday Discuss.* **2007**, *135*, 161–190.
- (26) Cardenas, C.; Ayers, P.; De Proft, F.; Tozer, D.; Geerlings, P. *Phys. Chem. Chem. Phys.* **2011**, *13*, 2285–2293.
- (27) Fukui, K. *Science* **1987**, *218*, 747–754.
- (28) Parr, R. G.; Yang, W. *J. Am. Chem. Soc.* **1984**, *106*, 4049–4050.
- (29) Yang, W. T.; Parr, R. G.; Pucci, R. *J. Chem. Phys.* **1984**, *81*, 2862–2863.
- (30) Ayers, P. W.; Levy, M. *Theor. Chem. Acc.* **2000**, *103*, 353–360.
- (31) Ayers, P. W.; Parr, R. G. *J. Am. Chem. Soc.* **2000**, *122*, 2010–2018.
- (32) Yang, W. T.; Parr, R. G. *Proc. Natl. Acad. Sci. U. S. A., Early Ed.* **1985**, *82*, 6723–6726.
- (33) Parr, R. G.; Szentpály, L. v.; Liu, S. *J. Am. Chem. Soc.* **1999**, *121*, 1922–1924.
- (34) Chattaraj, P. K.; Sarkar, U.; Roy, D. R. *Chem. Rev.* **2006**, *106*, 2065–2091.
- (35) Chattaraj, P. K.; Giri, S.; Duley, S. *J. Phys. Chem. Lett.* **2010**, *1*, 1064–1067.
- (36) Cardenas, C.; De Proft, F.; Chamorro, E.; Fuentealba, P.; Geerlings, P. *J. Chem. Phys.* **2008**, *128*, 034708.
- (37) Sablon, N.; Proft, F. D.; Geerlings, P. *J. Chem. Theory Comput.* **2009**, *5*, 1245–1253.
- (38) Tiznado, W.; Ona, O. B.; Bazterra, V. E.; Caputo, M. C.; Facelli, J. C.; Ferraro, M. B.; Fuentealba, P. *J. Chem. Phys.* **2005**, *123*, 214302.
- (39) Florez, E.; Tiznado, W.; Mondragón, F.; Fuentealba, P. *J. Phys. Chem. A* **2005**, *109*, 7815–7821.
- (40) Tiznado, W.; Oña, O. B.; Caputo, M. a. C.; Ferraro, M. B.; Fuentealba, P. *J. Chem. Theory Comput.* **2009**, *5*, 2265–2273.
- (41) Geerlings, P.; De Proft, F.; Langenaeker, W. *Chem. Rev.* **2003**, *103*, 1793–1873.
- (42) Chermette, H. *J. Comput. Chem.* **1999**, *20*, 129–154.
- (43) Perdew, J. P.; Parr, R. G.; Levy, M.; Balduz, J. L., Jr. *Phys. Rev. Lett.* **1982**, *49*, 1691–1694.
- (44) Yang, W. T.; Zhang, Y. K.; Ayers, P. W. *Phys. Rev. Lett.* **2000**, *84*, 5172–5175.
- (45) Ayers, P. W. *J. Math. Chem.* **2008**, *43*, 285–303.
- (46) Fuentealba, P.; Chamorro, E.; Cardenas, C. *Int. J. Quantum Chem.* **2007**, *107*, 37–45.
- (47) Ayers, P. W. *Phys. Chem. Chem. Phys.* **2006**, *8*, 3387–3390.
- (48) Bartolotti, L. J.; Ayers, P. W. *J. Phys. Chem. A* **2005**, *109*, 1146–1151.
- (49) Melin, J.; Ayers, P. W.; Ortiz, J. V. *J. Phys. Chem. A* **2007**, *111*, 10017–10019.
- (50) Cardenas, C.; Ayers, P. W.; Cedillo, A. *J. Chem. Phys.* **2011**, *134*, 174103–13.
- (51) Flores-Moreno, R. *J. Chem. Theory Comput.* **2009**, *6*, 48–54.
- (52) Martínez, J. *Chem. Phys. Lett.* **2009**, *478*, 310–322.
- (53) Ayers, P. W.; Parr, R. G.; Pearson, R. G. *J. Chem. Phys.* **2006**, *124*, 194107.
- (54) Bultinck, P.; Clarisse, D.; Ayers, P. W.; Carbo-Dorca, R. *Phys. Chem. Chem. Phys.* **2011**, *13*, 6110–6115.
- (55) Berkowitz, M. *J. Am. Chem. Soc.* **1987**, *109*, 4823–4825.
- (56) Fuentealba, P.; Florez, E.; Tiznado, W. *J. Chem. Theory Comput.* **2010**, *6*, 1470–1478.
- (57) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648–5652.
- (58) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785.
- (59) Krishnan, R.; Binkley, J. S.; Seeger, R.; Pople, J. A. *J. Chem. Phys.* **1980**, *72*, 650–654.
- (60) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, 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.; Peersson, 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.; Salvetti, O.; 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.

(61) Pople, J. A.; Headgordon, M.; Raghavachari, K. *J. Chem. Phys.* **1987**, *87*, 5968–5975.

(62) Kohout, M. *DGrid 4.4*, version 4.4 ed.; Radebeul, Germany, 2008; <http://www.cpfs.mpg.de/~kohout/dgrid.html>.

Assessment of Theoretical Methods for Complexes of Gold(I) and Gold(III) with Unsaturated Aliphatic Hydrocarbon: Which Density Functional Should We Choose?

Runhua Kang,^{†,§} Hui Chen,^{*,†} Sason Shaik,[‡] and Jiannian Yao^{*,†}

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

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

[§]Graduate University, Chinese Academy of Sciences, Beijing, 100049, China

 Supporting Information

ABSTRACT: Gold–substrate interaction is essential in gold-catalyzed organic transformations. This study uses high-level coupled cluster calculations with core–valence correlation and complete basis set (CBS) limit extrapolation as a reference, for assessing the performance of popular density functional theory (DFT) approximations for a variety of Au(I)/Au(III) complexes with unsaturated aliphatic hydrocarbon C_nH_m substrates (ethene, ethyne, and allene). The tested functionals cover from LDA to GGA and meta-GGA, and to hybrids and double hybrids (LSDA, PBE, M06-L, TPSS, B3LYP, PBE0, M06, M06-2X, TPSSh, B2-PLYP, B2GP-PLYP). Both the geometry and bond dissociation energy (D_e) of the Au– C_nH_m complexes are studied. Our findings show that B2GP-PLYP, PBE0, and B2-PLYP are the best performing functionals for this set of Au– C_nH_m complexes. DFT dispersion correction (DFT-D3), though very helpful for some functionals (e.g., B3LYP and B2-PLYP), does not uniformly improve the results of all functionals. Ab initio methods like MP2 and SCSMP2 are also tested. MP2 is found to be the worst performing method, and while SCSMP2 greatly improves the results, still its accuracy is lower than that of the best functionals, B2GP-PLYP, PBE0, and B2-PLYP.

1. INTRODUCTION

There has been a surge of interest in recent years in gold-catalyzed organic reactions with the concomitant rapid development and discovery of many new transformations in the field.^{1–11} The catalyzed reactions often involve unsaturated substrates with C=C, C=C=C, and C≡C functional groups. It is thought that the Au coordination to these functional groups is a key step in activating the substrates during these gold-catalyzed transformations.^{2,12} Due to the importance of these Au–substrate interactions, many theoretical and computational studies have been carried out to explore the interactions and bonding in Au–alkene^{13–21} and Au–alkyne complexes.^{19–24}

Density functional theory (DFT), which has had significant success in the transition metal chemistry area,^{25–38} has also been widely used for exploring many Au-catalyzed reaction mechanisms.^{39–71} However, what is still missing is a systematic study that assesses the performance of various approximate density functionals in reproducing the geometries and interactions between Au^I/Au^{III} and unsaturated aliphatic hydrocarbons. In this study, we employ coupled cluster CCSD(T) computations with an aim of achieving such a systematic assessment of the widely used DFT and some ab initio methods (e.g., MP2 and its variants) in reproducing geometries and bond dissociation energies (D_e) of Au–substrate complexes. As shown in Scheme 1, the target molecules AuL_x– C_nH_m used in the study are unsaturated substrates (C_nH_m) including ethene, ethyne, and allene, in

combination with a variety of Au^I/Au^{III}-ligated species related to real catalysts ($C_nH_m = C_2H_2, C_2H_4, C_3H_4$; AuL_x = Au^I, Au^ICl, Au^{III}Cl₃, Au^INHC, Au^IPH₃).

The coupled cluster benchmark data will rely on calibrated complete basis set (CBS) limit extrapolations including core–valence correlation effects, which is probably the highest ab initio level to have ever been applied to these Au– C_nH_m complexes. Useful knowledge about both the Au– C_nH_m interaction and its ab initio treatment is gained in this study, which would be helpful to future DFT and ab initio studies of Au-catalyzed organic transformations.

2. COMPUTATIONAL DETAILS

Treatment of Scalar Relativistic Effect. To take the scalar relativistic effects into account, we employed the new Stuttgart/Köln small-core relativistic pseudo-potential (PP)⁷² for gold, for both DFT and ab initio computations. To test whether such ECP treatment of scalar relativistic effects is sufficient, we performed MP2 calculations on naked Au(I) complexes using Stuttgart/Köln PP and compared the results to the calculations using the full electron third-order Douglas–Kroll–Hess (DKH) Hamiltonian.^{73,74} The DKH and PP calculations both included the Au-5s5p core–valence

Received: September 19, 2011

Published: November 11, 2011

Scheme 1. The Complexes Studied in This Work

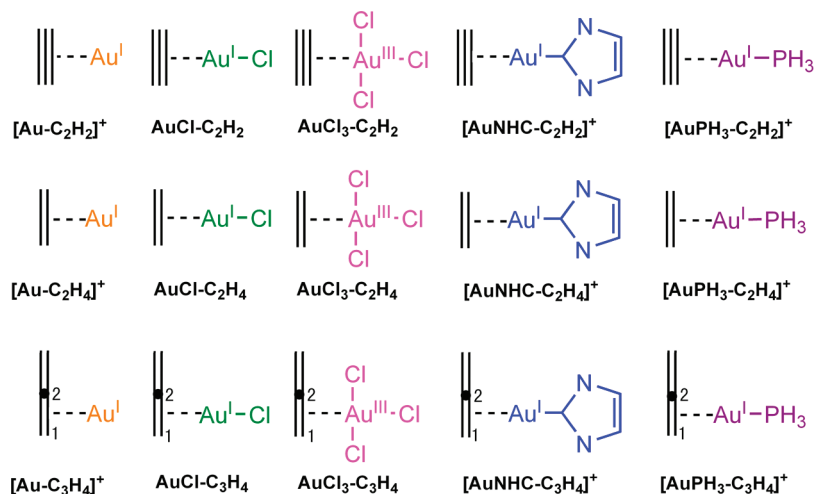


Table 1. Various Basis Sets and Their Abbreviations in This Work

elements	basis set ^a	abbreviation ^a	ref
Au	(aug-)cc-pVXZ-PP	(A)XZ	75
	(aug-)cc-pwCVXZ-PP	(A)wXZ	75
	(aug-)cc-pwCVTZ-DK	(A)wTZ-DK	75
C,H	(aug-)cc-pVTZ-DK	(A)TZ-DK	77
	(aug-)cc-pVXZ	(A)XZ	76
Cl,P	(aug-)cc-pV(X+d)Z	(A)XZ	78
C,N,Cl,P	cc-pwCVXZ	wXZ	76, 79

^a X represents the cardinal number of the correlation consistent basis set and could be D, T, Q, or 5 in this work.

correlation effect (note that the Au 5s orbital is lower in energy than the 4f orbitals in the DKH scheme at the Hartree–Fock level). The Au 5s5p core–valence correlation in the PP and DKH schemes used the corresponding correlation consistent weighted core–valence triple- ζ basis sets of (aug-)cc-pwCVTZ-PP and (aug-)cc-pwCVTZ-DK.⁷⁵ For the remaining main group elements (C, H), we adopted the (aug-)cc-pVTZ basis set⁷⁶ in the PP scheme and (aug-)cc-pVTZ-DK⁷⁷ in the DKH scheme. The results (see Tables S1 and S2 in Supporting Information (SI)) show that these two approaches generate similar geometries (maximal Au–C bond distance deviation of 0.002 Å) and D_e 's (maximal deviation of 0.96 kcal/mol). Considering the fact that the scope of this work is mainly the comparative study of different quantum chemical electronic structure methods, we can safely rely on the PP adopted for scalar relativistic effect. Because many basis sets are used in this work, we summarized their designations and our corresponding abbreviations in Table 1. “A” represents prefix “aug-” for a basis augmented by diffuse functions. “w” denotes the weighted core–valence correlated basis set. “X” is the cardinal number of the correlation consistent basis set, which could be D, T, Q, or 5. For third row p elements Cl and P, we used the “(X+d)” basis set that includes an additional set of d orbitals to improve the basis set extrapolation behavior. For clarity, we do not use the suffix “-PP” in our abbreviation for the basis set on the Au atom.

Density Functional Methods. All DFT calculations were done using the Gaussian 09⁸⁰ program. We employed 11 functionals which cover the range from LDA to (hybrid) GGA, (hybrid) meta-GGA, and double-hybrid functionals: LSDA,⁸¹ PBE,⁸² M06-L,⁸³ TPSS,⁸⁴ B3LYP,^{85–87} PBE0,^{82,88,89} M06,^{83,90} M06-2X,^{83,90} TPSSH,⁸⁴ B2-PLYP,⁹¹ and B2GP-PLYP.⁹²

Geometric performance tests were carried out only for the smaller naked Au(I) complexes $[\text{Au}^I-\text{C}_n\text{H}_m]^+$ ($\text{C}_n\text{H}_m = \text{C}_2\text{H}_2$, C_2H_4 , and C_3H_4), for which we could perform geometry optimization at the CCSD(T)/CBS level including Au 5s5p correlation. On the basis of the previous assessment of DFT optimized geometries of 5d transition-metal-containing complexes,⁹³ we selected five functionals, i.e., LSDA, B3LYP, PBE0, M06, and B2-PLYP, to test the optimized geometries, employing DZ and TZ basis sets.^{75,76} In double hybrid B2-PLYP geometry optimizations, the 1s electrons of C atoms were not correlated in MP2-type correlation treatments. In addition to these geometry tests, we also performed geometry optimizations for all complexes in Scheme 1 at the B3LYP/DZ level, which can be compared with the CCSD/DZ results (shown in Table S8 in the SI), and thereby provide a more general comparison between these two levels.

A test of D_e values obtained with B3LYP/AQZ^{75,76} showed a difference of up to 0.3 kcal/mol compared with the B3LYP/ATZ (see Table S3 in SI), hence indicating that the ATZ basis set gives values close to convergence for DFT. Therefore, all D_e calculations were done for all complexes in Scheme 1, using 10 of the functionals (excluding LSDA), in combination with the ATZ basis set.^{75,76} In addition, we tested the effect of dispersion correction using Grimme's DFT-D3 method,⁹⁴ and we employed two different short-range damping schemes, the original DFT-D3(0)^{94,95} and the Becke–Johnson scheme, DFT-D3-(BJ).^{96–100} We also tested for basis set superposition error (BSSE) correction for DFT using the B3LYP functional, and the results (Table S9 in the SI) indicate that the BSSE corrections are very small and thus can be neglected.

Ab Initio Methods. Four ab initio methods were employed using the Molpro2010 package.¹⁰¹ Two of these are coupled cluster methods (CCSD and CCSD(T)),^{102–104} and two are the perturbative methods, MP2¹⁰⁵ and SCMP2.¹⁰⁶ Basis set convergence in ab initio methods is not as fast as in DFT, and it

Table 2. $D_e/\text{CBS}_{\text{val}}$ (Valence Correlated CBS Limit, in kcal/mol) of $\text{AuCl}-\text{C}_2\text{H}_2$ Obtained by Extrapolation with Different Two-Point Extrapolation Formulas

D_e	eqs 3 and 4		eqs 1 and 2		
	CBS (ADZ-ATZ)	CBS (ATZ-AQZ)	CBS (ADZ-ATZ)	CBS (ATZ-AQZ)	CBS (AQZ-ASZ)
$\text{AuCl}-\text{C}_2\text{H}_2$	41.66	41.76	41.47	41.77	41.82
D_e	CBS (DZ-TZ)	CBS (TZ-QZ)	CBS (DZ-TZ)	CBS (TZ-QZ)	CBS (QZ-SZ)
$\text{AuCl}-\text{C}_2\text{H}_2$	40.04	41.35	39.78	41.37	41.82

requires extrapolation to the complete basis set (CBS) limit. We therefore selected the D_e datum for $\text{AuCl}-\text{C}_2\text{H}_2$ for calibration and performed extensive CBS limit extrapolation using a series of valence correlated basis sets from (A)DZ through (A)TZ and (A)QZ all the way to (A)SZ.^{75,76} Thus, we obtained a series of two-point extrapolation data from consecutive pairs of basis set extrapolation, denoted as (A)DZ-(A)TZ, (A)TZ-(A)QZ, and (A)QZ-(A)SZ. For two-point CBS limit extrapolations of Hartree-Fork (HF) and correlation energy, we tested a few proposed formulas. The first one is based on Helgaker's inverse cubic scaling dependence of the correlation energy on the basis set,¹⁰⁷ shown in eq 1.

$$E_{\text{corr},n} = E_{\text{corr,CBS}} + \frac{A}{n^3} \quad (1)$$

wherein two parameters A and $E_{\text{corr,CBS}}$ are to be determined by calculations with two basis sets, like (A)DZ-(A)TZ, while n is the cardinal number of the basis sets ($n = 2/3/4/5$ for (A)DZ/(A)TZ/(A)QZ/(A)SZ).

For the HF CBS limit extrapolation, a similar fifth power scaling dependence in eq 2¹⁰⁸

$$E_{\text{HF},n} = E_{\text{HF,CBS}} + \frac{A}{n^5} \quad (2)$$

was used in combination with eq 1

It was found before by Truhlar¹⁰⁹ that for CBS limit extrapolation involving a small basis like (A)DZ, which is the only affordable choice for the large complexes in this study, higher accuracy can be achieved by using the more general eq 3:

$$E_{\text{corr},n} = E_{\text{corr,CBS}} + \frac{A}{n^\beta} \quad (3)$$

The predetermined optimal parameter β is available for ADZ-ATZ ($\beta = 2.51$), DZ-TZ ($\beta = 2.46$), and (A)TZ-(A)QZ ($\beta = 3.05$) extrapolation.¹¹⁰ In combination with eq 3, we used the exponential scaling for HF CBS limit extrapolation,^{111,112} shown in eq 4

$$E_{\text{HF},n} = E_{\text{HF,CBS}} + A \exp(-\alpha\sqrt{n}) \quad (4)$$

wherein predetermined optimal parameter α is available for ADZ-ATZ ($\alpha = 4.30$), DZ-TZ ($\alpha = 4.42$), ATZ-AQZ ($\alpha = 5.79$), and TZ-QZ ($\alpha = 5.46$) CBS limit extrapolation.¹¹⁰

As summarized in Table 2, taking (A)QZ-(A)SZ extrapolation from eqs 1 and 2 as a reference, we find that (a) for the (A)DZ-(A)TZ extrapolation, eqs 3 and 4 produce better CBS limits for D_e than the values obtained from eqs 1 and 2 and (b) for the (A)TZ-(A)QZ extrapolation, eqs 3 and 4 generate almost the same CBS limit of D_e as that from eqs 1 and 2, which is a conceivable result since the optimal $\beta = 3.05$ in eq 3 for (A)TZ-(A)QZ extrapolation is very close to the theoretical

value of 3.0 in eq 1, and the difference of extrapolated absolute HF energies between eqs 4 and 2 with the (A)TZ-(A)QZ basis is already very small ($<0.4 \text{ mE}_h$). Thus, in this work, we used eqs 3 and 4 for (A)DZ-(A)TZ and (A)TZ-(A)QZ extrapolations, while for (A)QZ-(A)SZ extrapolation, we used eqs 1 and 2.

The above CBS limit extrapolation procedures are usable for CCSD(T) calculations. Thus, for MP2 and SCSMP2, only (A)DZ-(A)TZ extrapolations were done, employing eqs 3 and 4 with an optimal parameter β for MP2 ($\beta = 2.2$), as determined by Truhlar.¹⁰⁹

The highest level ab initio geometry optimizations for naked Au(I) complexes $[\text{Au}^1-\text{C}_n\text{H}_m]^+$ ($\text{C}_n\text{H}_m = \text{C}_2\text{H}_2, \text{C}_2\text{H}_4$, and C_3H_4) were performed with CCSD(T), MP2, and SCSMP2 with the wDZ/DZ and wTZ/TZ bases on Au/C,H atoms. The calculations include the Au 5s5p core-valence correlation. CCSD(T), MP2, and SCSMP2 CBS limit extrapolation was done then using eq 3 with the wDZ/DZ-wTZ/TZ basis set pair, as described above for the DZ-TZ basis set.

D_e Calculations. The geometries are uniformly optimized at the CCSD/DZ level for all of the complexes in Scheme 1 and used routinely for D_e calculations. Three ab initio methods, CCSD(T), MP2, and SCSMP2, and 10 functionals were employed in the D_e calculations for all of the complexes in Scheme 1.

The ab initio values were corrected to produce the final $D_e/\text{CBS}_{\text{final}}$ data based on eq 5:

$$D_e/\text{CBS}_{\text{final}} = D_e/\text{CBS}_{\text{val}} + \Delta D_e/\text{CBS}(5s5p) + \Delta D_e/\text{CBS}(CV) \quad (5)$$

Here, $D_e/\text{CBS}_{\text{val}}$ is the D_e value resulting from the ADZ-ATZ valence correlated CBS limit using eqs 3 and 4. $\Delta D_e/\text{CBS}(5s5p)$ is the Au 5s5p core-valence correlation correction for D_e from the wDZ/DZ-wTZ/TZ (in the order, Au/Cl,P,C,N,H) CBS limit extrapolation using eq 3. Lastly, $\Delta D_e/\text{CBS}(CV)$ is the core-valence correlation correction from other non-hydrogen atoms for D_e calculated from the wDZ/DZ-wTZ/TZ (in the order, Au,C,N,Cl,P/H) CBS limit extrapolation using eq 3, and excluding the correlation of the 1s electrons for Cl and P. All of the core-valence correlation corrections were calculated by the difference of two single point calculations, with and without the corresponding core-valence correlation using the same (A)wXZ basis set. Three ab initio methods, CCSD(T), MP2, and SCSMP2, were employed in D_e calculations for all of the complexes in Scheme 1. Ten functionals with the ATZ basis set were tested on the DFT side, as described above in the DFT subsection. In single point calculations with double-hybrid functionals, the core electrons of non-hydrogen atoms were correlated in an MP2-type correlation treatment.

The CBS D_e calculations were also corrected for basis set superposition error (BSSE) using the standard counterpoise

Table 3. D_e (kcal/mol) of AuCl–C₂H₂ Calculated at the CCSD(T) Level with Various Valence Correlated Basis Sets and CBS Limit Extrapolations, with and without BSSE Correction

D_e	ADZ	ATZ	AQZ	ASZ	CBS _{val} ^a (ADZ-ATZ)	CBS _{val} ^a (ATZ-AQZ)	CBS _{val} ^b (AQZ-ASZ)
without BSSE correction	39.48	40.87	41.40	41.61	41.66	41.76	41.82
with BSSE correction	35.35	39.11	40.69	41.22	41.13	41.73	41.75

^a CBS was obtained from eqs 3 and 4. ^b CBS was obtained from eqs 1 and 2.

Table 4. Key Bond Lengths (in Å) of Naked Au(I) Coordinate Complexes Optimized with Different Basis Sets at CCSD and CCSD(T) Levels

complex	parameter	CCSD/DZ	CCSD/ADZ	CCSD/TZ	CCSD/wDZ	CCSD(T)/DZ	CCSD(T)/CBS ^a (wDZ-wTZ)
[Au–C ₂ H ₂] ⁺	$r_{\text{Au}-\text{C}}$	2.191	2.191	2.179	2.168	2.175	2.135
	$r_{\text{C}-\text{C}}$	1.253	1.254	1.235	1.256	1.262	1.236
[Au–C ₂ H ₄] ⁺	$r_{\text{Au}-\text{C}}$	2.201	2.207	2.189	2.182	2.188	2.151
	$r_{\text{C}-\text{C}}$	1.404	1.408	1.392	1.408	1.413	1.399
[Au–C ₃ H ₄] ⁺	$r_{\text{Au}-\text{C}_1}$	2.157	2.164	2.146	2.142	2.149	2.117
	$r_{\text{Au}-\text{C}_2}$	2.217	2.221	2.212	2.188	2.190	2.158
	$r_{\text{C}_1-\text{C}_2}$	1.388	1.391	1.374	1.392	1.397	1.380

^a CCSD(T)/CBS was obtained by the extrapolation with eq 3 using wDZ/DZ-wTZ/TZ basis sets on Au/C,H to include the Au 5s5p correlation effect.

approach.^{113,114} The results for the AuCl–C₂H₂ complex shown in Table 3 indicate that BSSE corrected D_e values are consistently smaller than the corresponding ones without BSSE correction. As expected, with the larger basis sets like ATZ–AQZ or AQZ–ASZ, the BSSE corrections for the CBS limit are very small. While at the ADZ–ATZ CBS limit level, which is the only practical level for some relatively large complexes we studied here, and taking the (A)QZ–(A)SZ CBS limit extrapolation value as the reference, the calculations without BSSE correction generate much closer results than those with BSSE correction, in accord with a previous study of BSSE effect on interaction energy.¹¹⁵ This is a reflection that BSSE and basis set incompleteness error (BSIE) are oppositely signed for D_e . Thus, a favorable error cancellation occurs without BSSE correction. Therefore, in what follows, we show ADZ–ATZ CBS limit results without BSSE correction, while the BSSE corrected data are relegated to the Supporting Information (see Table S5 in the SI).

3. RESULTS AND DISCUSSION

3.1. Molecular Structures. *Naked Au(I) Coordinate Complexes.* To explore the factors that affect the geometries of Au–alkene/alkyne/allyene complexes, we first optimized the [Au^I–C_nH_m]⁺ (C_nH_m = C₂H₂, C₂H₄, and C₃H₄) complexes at various CCSD/DZ levels. To address the geometric influence of basis-set augmentation, we tested basis sets with diffuse functions (ADZ), a valence triple- ζ basis (TZ), and wDZ that considers the Au 5s5p correlation effects. In addition, CCSD(T) was also employed in geometry optimization combined with DZ to check the effect of perturbative triples (T) in the coupled cluster method. Key geometric parameters in these naked Au(I) coordinate complexes were collected in Table 4.

Basis Set and Au 5s5p Correlation Effect in Coupled Cluster Calculations. Inspection of the Au–C bond lengths in the three [Au^I–C_nH_m]⁺ complexes in Table 4 shows that augmenting the DZ basis set in CCSD with diffuse basis functions, from DZ to ADZ, results in minor Au–C bond lengthening (by 0.007 Å at most). However, improvement of the valence part, from DZ to

TZ, changes the bond lengths more significantly, e.g., shortening by up to 0.012 Å. Inclusion of the 5s5p electron correlation effect for gold, with the wDZ basis set, leads to maximal Au–C shortening, by 0.029 Å. Adding the triples (T) in CCSD(T)/DZ shortens the Au–C bond lengths by up to 0.027 Å compared with CCSD/DZ.

The above tests demonstrate that the quality of valence correlated basis set and the addition of Au 5s5p core–valence correlation and triples (T) all impart significant effects on the coupled cluster optimized geometries, while adding diffuse basis functions has a minor effect. Thus, we performed CCSD(T)/CBS calculations with the wDZ–wTZ basis set including the Au 5s5p correlation to obtain the most accurate and reliable geometric parameters of these [Au^I–C_nH_m]⁺ complexes, as shown for reference in the last column of Table 4. Comparing CCSD(T)/CBS Au–C bond lengths with CCSD/DZ ones, the former are shorter by about 0.040–0.059 Å than the latter.

Performance of DFT Methods. Having the reference geometric parameters for these [Au^I–C_nH_m]⁺ complexes, we can now assess the performance of DFT methods. To this end, we selected B3LYP, LSDA, M06, PBE0, and B2-PLYP to optimize the naked Au(I) complexes with DZ and TZ bases. The key bond lengths of these complexes optimized with TZ basis sets are collected in Table 5, while the results with DZ basis sets are reported in the Supporting Information (see Table S4 in the SI).

The best results using the CCSD(T)/CBS (wDZ-wTZ) reference are indicated in Table 5 by the bold font. Thus, PBE0 is the best method in the test set, with deviations of the Au–C bond lengths being only about –0.002 to +0.023 Å. This good performance of the PBE0 conforms to the previous geometry calibration of 5d transition metal complexes.⁹³ However, in contrast to the previous finding that LSDA is the best DFT in 5d transition metal complexes,⁹³ here, LSDA underestimates the Au–C bond lengths significantly (the deviations are about –0.062 to –0.077 Å). This conforms to the known tendency of LSDA to underestimate metal–ligand bond lengths in 3d and 4d transition metal species,^{116,117} a tendency that seems to persist here in the 5d transition metal complexes. B3LYP, which is the most popular functional in many areas,

Table 5. Key Bond Lengths (in Å) of Naked Au(I) Coordinate Complexes Optimized at Various DFT/TZ and ab Initio CBS Levels^a

		B3LYP/TZ	LSDA/TZ	M06/TZ	PBE0/TZ	B2-PLYP/TZ	SCSMP2/ CBS(wDZ-wTZ) ^b	MP2/CBS (wDZ-wTZ) ^b	CCSD(T)/ CBS(wDZ-wTZ) ^c
[Au–C ₂ H ₂] ⁺	<i>r</i> _{Au–C}	2.177	2.058	2.204	2.133	2.149	2.109	2.069	2.135
	<i>r</i> _{C–C}	1.233	1.250	1.227	1.237	1.241	1.240	1.247	1.236
[Au–C ₂ H ₄] ⁺	<i>r</i> _{Au–C}	2.203	2.086	2.222	2.158	2.171	2.127	2.092	2.151
	<i>r</i> _{C–C}	1.396	1.408	1.383	1.396	1.401	1.404	1.410	1.399
[Au–C ₃ H ₄] ⁺	<i>r</i> _{Au–C1}	2.148	2.053	2.161	2.115	2.126	2.094	2.062	2.117
	<i>r</i> _{Au–C2}	2.246	2.096	2.303	2.181	2.186	2.113	2.073	2.158
	<i>r</i> _{C1–C2}	1.375	1.388	1.363	1.377	1.382	1.386	1.392	1.380

^aThe best bond distances compared with reference CCSD(T)/CBS values are denoted in bold. ^bThe CBS limit extrapolation was obtained according to eq 3 ($\beta = 2.2$), with the wDZ/DZ-wTZ/TZ basis on Au/C,H. ^cThe CBS limit extrapolation was obtained according to eq 3 ($\beta = 2.46$), with the wDZ/DZ-wTZ/TZ basis on Au/C,H.

Table 6. D_e (kcal/mol) of AuCl–C₂H₂ Calculated from CCSD(T) with Various Valence Correlated Basis Sets and CBS Limit Extrapolations

	DZ	TZ	QZ	SZ	CBS ^a (DZ-TZ)	CBS ^a (TZ-QZ)	CBS ^b (QZ-SZ)
D_e	37.74	39.29	40.51	41.17	40.04	41.35	41.82
	ADZ	ATZ	AQZ	ASZ	CBS ^a (ADZ-ATZ)	CBS ^a (ATZ-AQZ)	CBS ^b (AQZ-ASZ)
D_e	39.48	40.87	41.40	41.61	41.66	41.76	41.82

^aExtrapolating to the CBS limit according to eqs 3 and 4. ^bExtrapolating to the CBS limit according to eqs 1 and 2.

including gold catalysis, overestimates Au–C bond lengths by about 0.031–0.088 Å. The M06 functional also overestimates the Au–C bond lengths by about 0.044–0.145 Å. It is noteworthy that the double-hybrid functional B2-PLYP produces the second lowest deviation among all tested functionals, with Au–C bond lengths overestimated by about 0.009–0.028 Å. Considering the basis set dependence of the double-hybrid functional (due to MP2 correlation component) compared with other functionals, B2-PLYP still has the potential to reach higher accuracy with a larger basis set than the used TZ basis set (see Table S4 in the SI for the trend of geometric change with increasing bases).

Performance of MP2 and SCSMP2. Geometric parameters obtained from the MP2/CBS level and SCSMP2/CBS level for the same naked Au(I) coordinate complexes are listed in Table 5. Compared with the CCSD(T)/CBS (wDZ-wTZ) reference values in Table 5, all Au–C bond lengths obtained with both MP2 and SCSMP2 are consistently shorter. Deviations of the Au–C bond lengths in these complexes are about –0.055 to –0.085 Å at the MP2/CBS level, while at the SCSMP2/CBS level, the deviations are significantly smaller, –0.023 to –0.045 Å, indicating the improvement of SCSMP2 over MP2.

3.2. Bond Dissociation Energies (D_e). We begin to discuss our extensive calibrations on AuCl–C₂H₂, for assessment of the accuracy of our adopted CCSD(T)/CBS procedure; then we show the results for all of the complexes at various ab initio and DFT levels.

Calibration with the AuCl–C₂H₂ Complex. Convergence of Valence-Correlated Basis Set. Table 6 summarizes these D_e results of AuCl–C₂H₂, calculated with CCSD(T) using the DZ-TZ-QZ-SZ and ADZ-ATZ-AQZ-ASZ basis sets, along with the corresponding CBS limit extrapolations.

It is seen that when the basis set quality improves, generally the D_e value increases, indicating that BSIE leads to underestimation

of D_e . Taking the highest-level CBS(AQZ-ASZ) value as a reference, it is clear that CBS extrapolation is absolutely necessary to overcome the BSIE problem, and the extrapolation improves the result even when it is performed from the low-end basis sets, such that the ADZ-ATZ extrapolation is better than the expensive ASZ D_e without CBS extrapolation, a situation that was frequently encountered in many previous high-level ab initio studies on various systems.^{109,118,119} It is noteworthy that the D_e value resulting from CBS(ADZ-ATZ) extrapolation (41.66 kcal/mol) is only 0.16 kcal/mol away from the CBS-(AQZ-ASZ) reference (41.82 kcal/mol), indicating a satisfactory balance between computational accuracy and the cost of the former approach. The beneficial effect of diffuse basis functions on D_e decreases along with the increasing quality of the valence basis set (1.74–1.58–0.89–0.44 kcal/mol along DZ–TZ–QZ–SZ series), which is understandable since the basis is becoming more complete and, hence, the role of the diffuse basis vanishes. The highest CBS(QZ-SZ) and CBS(AQZ-ASZ) values are identical to the second digit, which is a convincing indication that basis set convergence is achieved to a satisfactory level in the CBS extrapolation. At the double- ζ /triple- ζ CBS level, the effect of diffuse functions is still significant (1.62 kcal/mol). One interesting point is that if we add the D_e improvement from DZ to CBS(DZ-TZ) to the ADZ value, we get $D_e = 41.78$ kcal/mol, which is very close to the CBS(ADZ-ATZ) value (41.66 kcal/mol). This means we can use CBS(DZ-TZ) and ADZ single point, all of which are from a smaller basis set compared with ATZ, to get similar quality results of the CBS(ADZ-ATZ) value with a larger basis set. This CBS procedure could be very useful in those cases where diffuse functions are necessary but the ATZ calculation for large systems is prohibitively expensive in CBS extrapolation procedures.

Au 5s5p Correlation. Except for the valence correlation, we also explored the effects of Au 5s5p core–valence correlation

Table 7. Au 5s5p Core–Valence Correlation Corrections to D_e (kcal/mol) of AuCl–C₂H₂ from CCSD(T) with Different Basis Sets^a and CBS Limit Extrapolations

ΔD_e (AwDZ)	ΔD_e (AwTZ)	ΔD_e (AwQZ)	ΔD_e /CBS ^b (AwDZ–AwTZ)	ΔD_e /CBS ^b (AwTZ–AwQZ)
3.08	2.22	1.90	1.73	1.67
ΔD_e (wDZ)	ΔD_e (wTZ)	ΔD_e (wQZ)	ΔD_e /CBS ^b (wDZ–wTZ)	ΔD_e /CBS ^b (wTZ–wQZ)
2.68	2.05	1.81	1.70	1.64

^a (A)wXZ represents (A)wXZ/(A)XZ basis set on Au/Cl,C,H. ^b Extrapolating to the CBS limit with eqs 3, 4.

Table 8. CCSD(T)/CBS Calculated D_e (kcal/mol) of All Complexes, and Correction Components from Au-5s5p and Other Core–Valence Correlation Calculations

complex	D_e /ADZ	D_e /ATZ	D_e /CBS _{val} (ADZ–ATZ) ^a	ΔD_e /CBS(5s5p) ^b	ΔD_e /CBS(CV) ^c	D_e /CBS _{final} (CBS _{val} +5s5p+CV)
[Au–C ₂ H ₂] ⁺	51.84	53.19	53.91	2.61	–0.32	56.20
[Au–C ₂ H ₄] ⁺	60.36	62.30	63.34	2.91	–0.30	65.95
[Au–C ₃ H ₄] ⁺	61.29	62.90	63.73	3.05	–0.34	66.44
AuCl–C ₂ H ₂	39.48	40.87	41.66	1.70	–0.14	43.22
AuCl–C ₂ H ₄	43.95	46.00	47.14	1.75	–0.11	48.78
AuCl–C ₃ H ₄	44.26	45.79	46.65	1.86	–0.16	48.35
AuCl ₃ –C ₂ H ₂	28.37	28.68	29.08	0.56	–0.11	29.53
AuCl ₃ –C ₂ H ₄	38.80	39.21	39.78	0.71	–0.11	40.38
AuCl ₃ –C ₃ H ₄	37.51	37.54	37.88	0.70	–0.13	38.45
[AuNHC–C ₂ H ₂] ⁺	41.65	42.77	43.46	1.21	–0.17	44.50
[AuNHC–C ₂ H ₄] ⁺	45.00	46.55	47.47	1.25	–0.15	48.57
[AuNHC–C ₃ H ₄] ⁺	45.36	46.58	47.31	1.30	–0.19	48.42
[AuPH ₃ –C ₂ H ₂] ⁺	41.60	42.71	43.37	1.22	–0.10	44.49
[AuPH ₃ –C ₂ H ₄] ⁺	45.79	47.14	47.95	1.28	–0.06	49.17
[AuPH ₃ –C ₃ H ₄] ⁺	46.14	47.37	48.09	1.32	–0.09	49.32

^a The CBS_{val} limit was extrapolated from the CCSD(T) valence correlation calculated with the ADZ and ATZ basis sets. ^b Correction of the Au 5s5p core–valence correlation from the CCSD(T) CBS extrapolation with the wDZ/DZ and wTZ/TZ basis sets on Au/Cl,P,C,H,N. ^c Correction of the core–valence correlation of non-hydrogen main group elements (C, N, Cl, P) obtained from CCSD(T) CBS extrapolation with wDZ/DZ and wTZ/TZ basis sets on Au,C,N,Cl,P/H.

systematically with wXZ/XZ (X = D, T, Q) basis sets for Au/Cl, C,H atoms, as well as corresponding diffuse function augmented basis sets. The calculated ΔD_e 's due to the Au 5s5p core–valence correlation with the CCSD(T) method are summarized in Table 7.

First, it is seen that the Au 5s5p core–valence correlation effect computed at all levels consistently increases the D_e values. Second, the Au 5s5p core–valence correlation correction to D_e decreases with the increase in basis set quality. This is in line with the previous computational observation on the Au 5s5p correlation effect.⁷⁵ Last, we note that even though the AwDZ basis set generates a somewhat different ΔD_e value compared with the wDZ basis set, the ΔD_e /CBS values of two basis-set series, with and without diffuse basis functions, are quite close (within 0.03 kcal/mol). This result demonstrates that it is not necessary to use a much larger AwXZ basis instead of a wXZ basis to compute the Au 5s5p core–valence correlation effect. In addition, the wDZ–wTZ CBS limit is just 0.06 kcal/mol away from the reference wTZ–wQZ CBS limit, indicating that the wDZ–wTZ CBS limit is already sufficiently accurate.

More Au–C₂H₂/C₂H₄/C₃H₄ Complexes. CCSD(T)/CBS Reference Results. The calibration of the AuCl–C₂H₂ complex led to an appropriate general scheme to calculate the D_e values for other complexes. Table 8 summarizes these D_e values, calculated at the CCSD(T)/CBS_{final} level according to eq 5, in which we also

added the non-Au core–valence correlation correction ΔD_e –(CV)/CBS with wDZ/DZ–wTZ/TZ (basis set on Au,C,N,Cl,P/H) CBS limit extrapolation to reach higher accuracy. Although the magnitudes of the ΔD_e (CV) contributions are small, applying CBS limit extrapolation for this correction can change its signs from all positive at the CCSD(T)/DZ level to all negative, as those in Table 8.

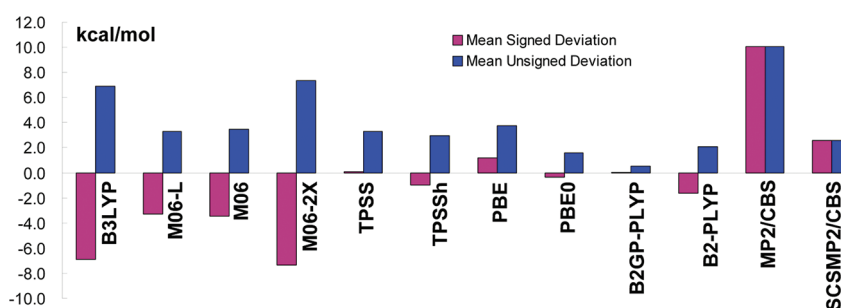
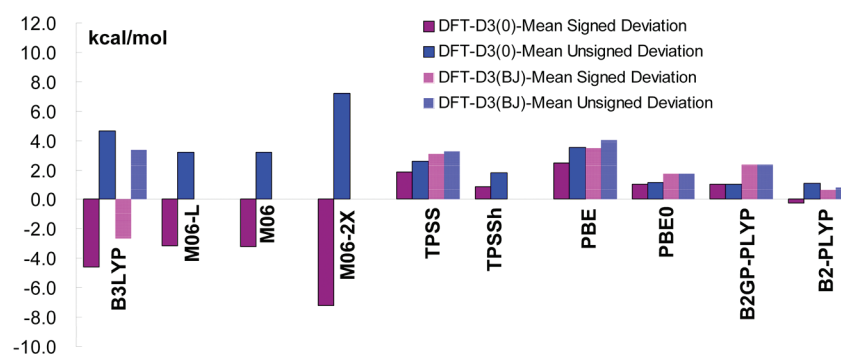
Inspecting the D_e /CBS(5s5p) values in Table 8, it is seen that the Au 5s5p core–valence correlation correction ΔD_e values for Au^{III}Cl₃–C_nH_m are smaller than for the other Au^IL_x–C_nH_m complexes. Just as found above for the group of naked Au^I complexes, the valence correlation extrapolation to the CBS limit is absolutely necessary because the improvements from the ATZ results to the CBS(ADZ–ATZ) limit are usually up to about 1 kcal/mol. Again, the Au^{III}Cl₃–C_nH_m complexes are different from other complexes due to their apparently smaller improvements (0.34–0.57 vs 0.66–1.14 kcal/mol) from the ATZ results to the ADZ–ATZ CBS limit.

Considering the CCSD(T)/CBS_{final} values, the D_e 's for the naked Au^I complexes [Au–C_nH_m]⁺ are the largest among the studied complexes, while the AuCl₃–C_nH_m species have the smallest D_e values. For the complexes with other ligands like AuCl–C_nH_m, [AuNHC–C_nH_m]⁺, and [AuPH₃–C_nH_m]⁺, the variations of the D_e values for the same unsaturated hydrocarbon substrate are no more than about 1 kcal/mol. Concerning the

Table 9. Mean Signed (S) and Unsigned (U) Deviations of Various DFT Methods without and with DFT-D3 Dispersion Corrections from CCSD(T)/CBS_{final}^a

		B3LYP	M06-L	M06	M06-2X	TPSS	TPSSh	PBE	PBE0	B2GP-PLYP	B2-PLYP
without DFT-D3	U	6.91	3.31	3.48	7.33	3.31	2.97	3.72	1.60	0.52	2.07
	S	-6.91	-3.31	-3.48	-7.33	0.10	-0.97	1.22	-0.36	0.05	-1.61
DFT-D3(0) ^b	U	4.71	3.21	3.25	7.25	2.56	1.85	3.52	1.14	0.99	1.08
	S	-4.61	-3.21	-3.25	-7.25	1.83	0.80	2.41	1.00	0.99	-0.30
DFT-D3(BJ) ^c	U	3.36	— ^d	— ^d	— ^d	3.22	— ^e	3.97	1.74	2.35	0.77
	S	-2.69	— ^d	— ^d	— ^d	3.04	— ^e	3.42	1.74	2.35	0.59

^a The smallest three mean unsigned deviations of each set are denoted in bold. ^b Using zero-damping in DFT-D3. ^c Using Becke–Johnson (BJ) damping in DFT-D3. ^d There are no BJ parameters for the M06 series of functionals. ^e Currently, the BJ parameters of TPSSh could have some problem, see ref 120; thus these data are omitted.

**Figure 1.** Mean signed and unsigned deviations of calculated D_e with various DFT/ATZ and ab initio (CBS_{final}) methods taking CCSD(T)/CBS_{final} as a reference.**Figure 2.** Mean signed and unsigned deviations of calculated D_e with various DFT-D3/ATZ methods taking CCSD(T)/CBS_{final} as a reference.

comparison between different C_nH_m 's within one AuL_x , the D_e of C_2H_2 is apparently and consistently smaller than those of C_2H_4 and C_3H_6 , which is in line with previous theoretical calculations favoring the Au–alkene over the Au–alkyne interaction thermodynamically,²⁰ while the difference of D_e between the latter two is usually quite small, except for $AuCl_3$ systems.

In summary of our ab initio treatment of D_e of Au complexes, we note the following:

- ADZ-ATZ CBS is necessary and sufficient for good accuracy (within about 0.2 kcal/mol from (A)QZ-(A)SZ CBS value). Note that diffuse functions are still important at this level (having about a 1–2 kcal/mol effect on D_e).
- Au-5s5p core–valence correlation is important (up to about 3 kcal/mol), and DZ-TZ CBS limit extrapolation is necessary to get the converged values. Here, diffuse functions are not important.

- Non-Au core–valence correlation is not very important (within -0.34 kcal/mol).

Performance of Various DFT Methods. The CCSD(T)/CBS_{final} values of D_e can serve as a benchmark for testing the various DFT functionals, with an aim of hopefully finding one or more that are uniformly better than others for these problems.

For the DFT data of each of the complexes, see Table S6 and S7 in the SI. Table 9 summarizes the statistical analysis of the mean signed/unsigned deviations of various DFT methods without and with dispersion corrections. These deviations are also shown schematically in Figures 1 and 2, where Figure 1 shows also the performance of two tested ab initio methods.

From the D_e values in Table 9 and Figure 1, we find that ranking of the tested functionals without dispersion corrections, in order of increasing mean unsigned deviation, is B2GP-PLYP < PBE0 < B2-PLYP < SCSPMP2 < TPSSh < TPSS \approx M06-L < M06 < PBE < B3LYP < M06-2X < MP2. So it can be seen that good

geometric performances of DFT methods correlate with their good energetic performances. In all of the tested methods, B2GP-PLYP provides the best results compared with CCSD(T)/CBS_{final}, with the maximal deviation of 1.14 kcal/mol, and a very small mean signed/unsigned deviation of 0.05/0.52 kcal/mol. For this set of Au complexes, B2GP-PLYP, which is a reparametrized B2-PLYP, outperforms its parent. The second best one is PBE0, which is best among all of the tested (hybrid)-(meta-)GGA functionals, with a maximal deviation of -3.79 kcal/mol and a small mean signed/unsigned deviation of -0.36/1.60 kcal/mol. The third best one is B2-PLYP, with a maximal deviation of -3.75 kcal/mol and a small mean signed/unsigned deviation of -1.61/2.07 kcal/mol. The popular B3LYP functional produces the second largest deviations in all tested DFT, with a maximal deviation of -12.49 kcal/mol and mean signed/unsigned deviations of -6.91/6.91 kcal/mol. It is interesting to note that B3LYP, M06-L, M06, and M06-2X consistently underestimate D_e (thus have signed/unsigned deviations of exactly same magnitude but opposite sign), while other tested functionals do not behave similarly. M06-2X performs much worse than the other two in the M06 series, which indicates that a high ratio of exact exchange is not beneficial in these Au complexes of late transition metals.

After adding the DFT-D3 dispersion correction, as shown in Table 9 and Figure 2, the deviations of some tested functionals were consistently reduced, with considerable improvement for B3LYP and B2-PLYP. As found before,^{92,121} the results of the M06 series functionals are almost not affected by adding dispersion corrections. For some tested density functionals, like TPSS, PBE, and PBE0, DFT-D3 corrections do not uniformly improve the results. DFT-D3(0) with zero short-range damping improves the results of TPSS, PBE, and PBE0, while DFT-D3(BJ) with BJ-damping does not. For B2GP-PLYP, both DFT-D3 corrections turn out to spoil an already good performance. After and before DFT-D3 corrections, the best three functionals remain to be PBE0, B2GP-PLYP, and B2P-PLYP, where the first rank depends on the dispersion correction scheme, B2GP-PLYP (no dispersion correction and DFT-D3(0) correction) or B2-PLYP (DFT-D3(BJ) correction).

Performance of MP2 and SCSMP2. As shown in Table 9 and Figure 1, the D_e values obtained at the MP2/CBS_{final} and SCSMP2/CBS_{final} levels are all overestimated compared to the reference values at the CCSD(T)/CBS_{final} level. Especially poor is the MP2/CBS method, which has the largest mean deviation (10.08 kcal/mol) and largest maximal deviation (14.41 kcal/mol) of D_e among all of the tested methods. SCSMP2 significantly reduced the mean deviation to about 2.6 kcal/mol, and the maximal deviation to about 5.3 kcal/mol, but its performance compared with many tested DFT methods is still not impressive, especially considering that it is more computationally costly than most DFT approaches.

4. CONCLUSIONS

We have systematically studied herein the geometries and bond dissociation energies (D_e) of typical AuL_x-C_nH_m complexes (Scheme 1) by employing ab initio coupled cluster, MP2, SCSMP2, and DFT (B3LYP, LSDA, M06, M06L, M06-2X, PBE, PBE0, TPSS, TPSSh, B2-PLYP, and B2GP-PLYP) calculations. Our reference CCSD(T) calculations include the calibrated valence correlation complete basis set (CBS) extrapolation and core-valence correlation corrections, which are shown to be

important for the achievement of high accuracy. For geometry optimization, the best tested DFT method is the hybrid GGA PBE0 followed by the double hybrid B2-PLYP. Unlike LSDA, the extraordinary performance of PBE0 in geometry conforms to the previous calibration on 5d transition metal complexes.⁹³ For D_e , the most accurate DFT methods are the two double-hybrid functionals B2-PLYP and B2GP-PLYP, with ranking order depending on whether one does or does not add dispersion correction and the particular scheme thereof (D3(0) or D3(BJ)). With or without DFT dispersion correction, the smallest mean unassigned deviations of these two double-hybrid methods are all within about 1 kcal/mol. The largest unassigned deviations for them are about 1–3 kcal/mol. Within the tested GGA or hybrid GGA, PBE0 is again the best one, with performance which is just slightly worse than that of the best of the two double hybrids with DFT-D3(0). Thus, the good geometric performance is found to correlate with good energetics of DFT methods in this study. B3LYP is significantly improved by dispersion corrections, but it is still outperformed by PBE0. DFT-D3 dispersion does not yield a uniform improvement for all of the functionals we tested here. For both geometry and D_e , MP2 significantly underestimates/directions of most tested DFT functionals like B3LYP. SCSMP2 greatly alleviates the poor performance of MP2, but it is not better than the best performing DFT methods for these Au complexes. All in all, the double-hybrid functionals tested in this study showed very promising performance on these Au(I)/Au(III) complexes with unsaturated molecules.

In general, our recommended cost-effective methods for Au-unsaturated hydrocarbon complexes are three functionals: two double hybrids, B2GP-PLYP and B2-PLYP, and one hybrid GGA, PBE0. It would be interesting to know whether these functionals can keep their good performance in theoretical studies of homogeneous gold-catalyzed reactions, and whether there exist alternative optimal functionals. These studies are underway in our group.

■ ASSOCIATED CONTENT

Supporting Information. Nine tables and Cartesian coordinates of the studied complexes used in D_e calculations. This material is available free of charge via the Internet at <http://pubs.acs.org>.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: chenh@iccas.ac.cn; jnyao@iccas.ac.cn.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENT

This work was supported by the Chinese Academy of Sciences and National Basic Research Program of China (No. 2011CB808402). The research of S.S. is supported by the Israel Science Foundation (ISF Grant 53/09). The authors are thankful to Stefan Grimme for helpful advice in DFT-D3 calculations.

■ REFERENCES

- (1) Hashmi, A. S. K. *Chem. Rev.* **2007**, *107*, 3180–3211.

- (2) Fürstner, A.; Davies, P. W. *Angew. Chem., Int. Ed.* **2007**, *46*, 3410–3449.
- (3) Li, Z.; Brouwer, C.; He, C. *Chem. Rev.* **2008**, *108*, 3239–3265.
- (4) Arcadi, A. *Chem. Rev.* **2008**, *108*, 3266–3325.
- (5) Jiménez-Núñez, E.; Echavarren, A. M. *Chem. Rev.* **2008**, *108*, 3326–3350.
- (6) Pina, C. D.; Falletta, E.; Prati, L.; Rossi, M. *Chem. Soc. Rev.* **2008**, *37*, 2077–2095.
- (7) Gorin, D. J.; Sherry, B. D.; Toste, F. D. *Chem. Rev.* **2008**, *108*, 3351–3378.
- (8) Krause, N.; Winter, C. *Chem. Rev.* **2011**, *111*, 1994–2009.
- (9) Bandini, M. *Chem. Soc. Rev.* **2011**, *40*, 1358–1367.
- (10) Boorman, T. C.; Larrosa, I. *Chem. Soc. Rev.* **2011**, *40*, 1910–1925.
- (11) Corma, A.; Leyva-Pérez, A.; Sabater, M. J. *Chem. Rev.* **2011**, *111*, 1657–1712.
- (12) Schmidbaur, H.; Schier, A. *Organometallics* **2010**, *29*, 2–23.
- (13) Nicolas, G.; Spiegelmann, F. *J. Am. Chem. Soc.* **1990**, *112*, 5410–5419.
- (14) Hertwig, R. H.; Koch, W.; Schröder, D.; Schwarz, H. *J. Phys. Chem.* **1996**, *100*, 12253–12260.
- (15) Mendizabal, F. *Int. J. Quantum Chem.* **1999**, *73*, 317–324.
- (16) Kim, C. K.; Lee, K. A.; Kim, C. K.; Lee, B. S.; Lee, H. W. *Chem. Phys. Lett.* **2004**, *391*, 321–324.
- (17) Stringer, K. L.; Citir, M.; Metz, R. B. *J. Phys. Chem. A* **2004**, *108*, 6996–7002.
- (18) Hertwig, R. H.; Hrušák, J.; Schröder, D.; Koch, W.; Schwarz, H. *Chem. Phys. Lett.* **1995**, *236*, 194–200.
- (19) Yamamoto, Y. *J. Org. Chem.* **2007**, *72*, 7817–7831.
- (20) Garcí-Mota, M.; Cabello, N.; Maseras, F.; Echavarren, A. M.; Pérez-Ramírez, J.; Lopez, N. *ChemPhysChem* **2008**, *9*, 1624–1629.
- (21) Nechaev, M. S.; Rayón, V. M.; Frenking, G. *J. Phys. Chem. A* **2004**, *108*, 3134–3142.
- (22) Flügge, S.; Anoop, A.; Goddard, R.; Thiel, W.; Fürstner, A. *Chem.—Eur. J.* **2009**, *15*, 8558–8565.
- (23) Zuccaccia, D.; Belpassi, L.; Rocchigiani, L.; Tarantelli, F.; Macchioni, A. *Inorg. Chem.* **2010**, *49*, 3080–3082.
- (24) Yamamoto, Y.; Gridnev, I. D.; Patil, N. T.; Jin, T. *Chem. Commun.* **2009**, 5075–5087.
- (25) Siegbahn, P. E. M. *Adv. Chem. Phys.* **1996**, *93*, 333–378.
- (26) Niu, S.; Hall, M. B. *Chem. Rev.* **2000**, *100*, 353–405.
- (27) Balcells, D.; Clot, E.; Eisenstein, O. *Chem. Rev.* **2010**, *110*, 749–823.
- (28) Siegbahn, P. E. M.; Blomberg, M. R. A. *Chem. Rev.* **2000**, *100*, 421–437.
- (29) Noodleman, L.; Lovell, T.; Han, W.-G.; Li, J.; Himo, F. *Chem. Rev.* **2004**, *104*, 459–508.
- (30) Baik, M.-H.; Newcomb, M.; Friesner, R. A.; Lippard, S. J. *Chem. Rev.* **2003**, *103*, 2385–2419.
- (31) Shaik, S.; Cohen, S.; Wang, Y.; Chen, H.; Kumar, D.; Thiel, W. *Chem. Rev.* **2010**, *110*, 949–1017.
- (32) Cramer, C. J.; Truhlar, D. G. *Phys. Chem. Chem. Phys.* **2009**, *11*, 10757–10816.
- (33) Furche, F.; Perdew, J. P. *J. Chem. Phys.* **2006**, *124*, 044103.
- (34) Jensen, K. P.; Roos, B. O.; Ryde, U. *J. Chem. Phys.* **2007**, *126*, 014103.
- (35) Quintal, M. M.; Karton, A.; Iron, M. A.; Boese, A. D.; Martin, J. M. L. *J. Phys. Chem. A* **2006**, *110*, 709–716.
- (36) Swart, M. J. *Chem. Theory Comput.* **2008**, *4*, 2057–2066.
- (37) Riley, K. E.; Merz, K. M., Jr. *J. Phys. Chem. A* **2007**, *111*, 6044–6053.
- (38) de Jong, G. T.; Bickelhaupt, F. M. *J. Chem. Theory Comput.* **2006**, *2*, 322–335.
- (39) Paton, R. S.; Maseras, F. *Org. Lett.* **2009**, *11*, 2237–2240.
- (40) Marion, N.; Lemièrre, G.; Correa, A. *Chem.—Eur. J.* **2009**, *15*, 3243–3260.
- (41) Nieto-Oberhuber, C.; Muñoz, M. P.; Buñuel, E.; Nevado, C.; Cárdenas, D. J.; Echavarren, A. M. *Angew. Chem., Int. Ed.* **2004**, *43*, 2402–2406.
- (42) Xia, Y.; Huang, G. *J. Org. Chem.* **2010**, *75*, 7842–7854.
- (43) Zhang, J.; Shen, W.; Li, L.; Li, M. *Organometallics* **2009**, *28*, 3129–3139.
- (44) Lee, Y. T.; Kang, Y. K.; Chung, Y. K. *J. Org. Chem.* **2009**, *74*, 7922–7934.
- (45) Shi, F. Q.; Li, X.; Xia, Y.; Zhang, L.; Yu, Z. X. *J. Am. Chem. Soc.* **2007**, *129*, 15503–15512.
- (46) Fang, R.; Yang, L.; Wang, Y. *Org. Biomol. Chem.* **2011**, *9*, 2760–2770.
- (47) Faza, O. N.; López, C. S.; Álvarez, R.; de Lera, A. R. *J. Am. Chem. Soc.* **2006**, *128*, 2434–2437.
- (48) Cheong, P. H.-Y.; Morganeli, P.; Luzung, M. R.; Houk, K. N.; Toste, F. D. *J. Am. Chem. Soc.* **2008**, *130*, 4517–4526.
- (49) Wang, M.-Z.; Zhou, C.-Y.; Guo, Z.; Wong, E. L.-M.; Wong, M.-K.; Che, C.-M. *Chem. Asian J.* **2011**, *6*, 812–824.
- (50) Jiang, M.; Liu, L.-P.; Shi, M.; Li, Y. *Org. Lett.* **2010**, *12*, 116–119.
- (51) Correa, A.; Marion, N.; Fensterbank, L.; Malacria, M.; Nolan, S. P.; Cavallo, L. *Angew. Chem., Int. Ed.* **2008**, *47*, 718–721.
- (52) Lemièrre, G.; Gandon, V.; Agenet, N.; Goddard, J.-P.; de Kozak, A.; Aubert, C.; Fensterbank, L.; Malacria, M. *Angew. Chem., Int. Ed.* **2006**, *45*, 7596–7599.
- (53) Lemièrre, G.; Gandon, V.; Cariou, K.; Hours, A.; Fukuyama, T.; Dhimane, A.-L.; Fensterbank, L.; Malacria, M. *J. Am. Chem. Soc.* **2009**, *131*, 2993–3006.
- (54) Liu, L.-P.; Hammond, G. B. *Chem. Asian J.* **2009**, *4*, 1230–1236.
- (55) Nieto-Oberhuber, C.; López, S.; Muñoz, M. P.; Cárdenas, D. J.; Buñuel, E.; Nevado, C.; Echavarren, A. M. *Angew. Chem., Int. Ed.* **2005**, *44*, 6146–6148.
- (56) Teles, J. H.; Brode, S.; Chabanas, M. *Angew. Chem., Int. Ed.* **1998**, *37*, 1415–1418.
- (57) Dudnik, A. S.; Xia, Y.; Li, Y.; Gevorgyan, V. *J. Am. Chem. Soc.* **2010**, *132*, 7645–7655.
- (58) Kovács, G.; Lledós, A.; Ujaque, G. *Organometallics* **2010**, *29*, 3252–3260.
- (59) Liu, Y.; Zhang, D.; Zhou, J.; Liu, C. *J. Phys. Chem. A* **2010**, *114*, 6164–6170.
- (60) Zhu, R.-X.; Zhang, D.-J.; Guo, J.-X.; Mu, J.-L.; Duan, C.-G.; Liu, C.-B. *J. Phys. Chem. A* **2010**, *114*, 4689–4696.
- (61) Mauleón, P.; Krinsky, J. L.; Toste, F. D. *J. Am. Chem. Soc.* **2009**, *131*, 4513–4520.
- (62) Sordo, T. L.; Ardura, D. *Eur. J. Org. Chem.* **2008**, 3004–3013.
- (63) Gourlaouen, C.; Marion, N.; Nolan, S. P.; Maseras, F. *Org. Lett.* **2009**, *11*, 81–84.
- (64) Straub, B. F. *Chem. Commun.* **2004**, 1726–1728.
- (65) Nun, P.; Gaillard, S.; Poater, A.; Cavallo, L.; Nolan, S. P. *Org. Biomol. Chem.* **2011**, *9*, 101–104.
- (66) Xia, Y.; Dudnik, A. S.; Li, Y.; Gevorgyan, V. *Org. Lett.* **2010**, *12*, 5538–5541.
- (67) Seraya, E.; Slack, E.; Ariafard, A.; Yates, B. F.; Hyland, C. J. T. *Org. Lett.* **2010**, *12*, 4768–4771.
- (68) Roithová, J.; Hrušák, J.; Schröder, D.; Schwarz, H. *Inorg. Chim. Acta* **2005**, *358*, 4287–4292.
- (69) Marion, N.; Carlqvist, P.; Gealageas, R.; de Frémont, P.; Maseras, F.; Nolan, S. P. *Chem.—Eur. J.* **2007**, *13*, 6437–6451.
- (70) Kovács, G.; Ujaque, G.; Lledós, A. *J. Am. Chem. Soc.* **2008**, *130*, 853–864.
- (71) Yang, L.; Fang, R.; Wang, Y. *Comput. Theor. Chem.* **2011**, *965*, 180–185.
- (72) Figgen, D.; Rauhut, G.; Dolg, M.; Stoll, H. *Chem. Phys.* **2005**, *311*, 227–244.
- (73) Reiher, M.; Wolf, A. *J. Chem. Phys.* **2004**, *121*, 10945–10956.
- (74) Reiher, M.; Wolf, A. *J. Chem. Phys.* **2004**, *121*, 2037–2047.
- (75) Peterson, K. A.; Puzzarini, C. *Theor. Chem. Acc.* **2005**, *114*, 283–296.
- (76) Dunning, T. H. J. *J. Chem. Phys.* **1989**, *90*, 1007–1023.
- (77) de Jong, W. A.; Harrison, R. J.; Dixon, D. A. *J. Chem. Phys.* **2001**, *114*, 48–53.
- (78) Dunning, T. H. J.; Peterson, K. A.; Wilson, A. K. *J. Chem. Phys.* **2001**, *114*, 9244–9253.

- (79) Woon, D. E.; Dunning, T. H. *J. Chem. Phys.* **1993**, *98*, 1358–1371.
- (80) 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.; Comperts, 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.
- (81) Vosko, S. H.; Wilk, L.; Nusair, M. *Can. J. Phys.* **1980**, *58*, 1200–1211.
- (82) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865–3868.
- (83) Zhao, Y.; Truhlar, D. G. *J. Chem. Phys.* **2006**, *125*, 194101.
- (84) Tao, J.; Perdew, J. P.; Staroverov, V. N.; Scuseria, G. E. *Phys. Rev. Lett.* **2003**, *91*, 146401.
- (85) Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098–3100.
- (86) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785–789.
- (87) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648–5652.
- (88) Ernzerhof, M.; Scuseria, G. E. *J. Chem. Phys.* **1999**, *110*, 5029–5036.
- (89) Adamo, C.; Barone, V. *J. Chem. Phys.* **1999**, *110*, 6158–6170.
- (90) Zhao, Y.; Truhlar, D. G. *Theor. Chem. Acc.* **2008**, *120*, 215–241.
- (91) Grimme, S. *J. Chem. Phys.* **2006**, *124*, 034108.
- (92) Karton, A.; Tarnopolsky, A.; Lamere, J.-F.; Schatz, G. C.; Martin, J. M. L. *J. Phys. Chem. A* **2008**, *112*, 12868–12886.
- (93) Bühl, M.; Reimann, C.; Pantazis, D. A.; Bredow, T.; Neese, F. *J. Chem. Theory Comput.* **2008**, *4*, 1449–1459.
- (94) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. *J. Chem. Phys.* **2010**, *132*, 154104.
- (95) Goerigk, L.; Grimme, S. *J. Chem. Theory Comput.* **2011**, *7*, 291–309.
- (96) Becke, A. D.; Johnson, E. R. *J. Chem. Phys.* **2005**, *122*, 154101.
- (97) Johnson, E. R.; Becke, A. D. *J. Chem. Phys.* **2005**, *123*, 024101.
- (98) Johnson, E. R.; Becke, A. D. *J. Chem. Phys.* **2006**, *124*, 174104.
- (99) Grimme, S.; Ehrlich, S.; Goerigk, L. *J. Comput. Chem.* **2011**, *32*, 1456–1465.
- (100) Goerigk, L.; Grimme, S. *Phys. Chem. Chem. Phys.* **2011**, *13*, 6670–6688.
- (101) Werner, H.-J.; Knowles, P. J.; Lindh, R.; Manby, F. R.; Schütz, M.; Celani, P.; Korona, T.; Mitrushenkov, A.; Rauhut, G.; Adler, T. B.; Amos, R. D.; Bernhardsson, A.; Berning, A.; Cooper, D. L.; Deegan, M. J. O.; Dobbyn, A. J.; Eckert, F.; Goll, E.; Hampel, C.; Hetzer, G.; Hrenar, T.; Knizia, G.; Köppl, C.; Liu, Y.; Lloyd, A. W.; Mata, R. A.; May, A. J.; McNicholas, S. J.; Meyer, W.; Mura, M. E.; Nicklass, A.; Palmieri, P.; Pflüger, K.; Pitzer, R.; Reiher, M.; Schumann, U.; Stoll, H.; Stone, A. J.; Tarroni, R.; Thorsteinsson, T.; Wang, M.; Wolf, A. MOLPRO, version 2010.1, a package of ab initio programs. See <http://www.molpro.net> (accessed Nov. 2011).
- (102) Knowles, P. J.; Hampel, C.; Werner, H.-J. *J. Chem. Phys.* **1993**, *99*, 5219–5227.
- (103) Watts, J. D.; Gauss, J.; Bartlett, R. J. *J. Chem. Phys.* **1993**, *98*, 8718–8733.
- (104) Purvis, G. D., III; Bartlett, R. J. *J. Chem. Phys.* **1982**, *76*, 1910–1918.
- (105) Møller, C.; Plesset, M. S. *Phys. Rev.* **1934**, *46*, 618–622.
- (106) Grimme, S. *J. Chem. Phys.* **2003**, *118*, 9095–9102.
- (107) Helgaker, T.; Klopper, W.; Koch, H.; Noga, J. *J. Chem. Phys.* **1997**, *106*, 9639–9646.
- (108) Parthiban, S.; Martin, J. M. L. *J. Chem. Phys.* **2001**, *114*, 6014–6029.
- (109) Truhlar, D. G. *Chem. Phys. Lett.* **1998**, *294*, 45–48.
- (110) Neese, F.; Valeev, E. F. *J. Chem. Theory Comput.* **2011**, *7*, 33–43.
- (111) Zhong, S. J.; Barnes, E. C.; Petersson, G. A. *J. Chem. Phys.* **2008**, *129*, 184116.
- (112) Karton, A.; Martin, J. M. L. *Theor. Chem. Acc.* **2006**, *115*, 330–333.
- (113) Cramer, C. J. *Essentials of Computational Chemistry: Theories and Models*, 2nd ed.; John Wiley & Sons Inc.: Chichester, West Sussex, U.K., 2004; pp 195–196.
- (114) Boys, S. F.; Bernardi, F. *Mol. Phys.* **1970**, *19*, 553–566.
- (115) Masamura, M. *Theor. Chem. Acc.* **2001**, *106*, 301–313.
- (116) Bühl, M.; Kabrede, H. *J. Chem. Theory Comput.* **2006**, *2*, 1282–1290.
- (117) Waller, M. P.; Braun, H.; Hojdis, N.; Bühl, M. *J. Chem. Theory Comput.* **2007**, *3*, 2234–2242.
- (118) Helgaker, T.; Klopper, W.; Halkier, A.; Bak, K. L.; Jørgensen, P.; Olsen, J. In *Quantum-Mechanical Prediction of Thermochemical Data*; Cioslowski, J., Ed.; Kluwer Academic: Dordrecht, The Netherlands, 2001; pp 1–30.
- (119) Korth, M.; Grimme, S. *J. Chem. Theory Comput.* **2009**, *4*, 993–1003.
- (120) <http://toc.uni-muenster.de/DFTD3/index.html> (accessed Sep. 13, 2011).
- (121) Kozuch, S.; Gruzman, D.; Martin, J. M. L. *J. Phys. Chem. C* **2010**, *114*, 20801–20808.

On the Importance of the Orbital Relaxation in Ground-State Coupled Cluster Calculations in Solution with the Polarizable Continuum Model of Solvation

Marco Caricato* and Giovanni Scalmani

Gaussian, Inc., 340 Quinipiac Street, Building 40, Wallingford, Connecticut 06492, United States

ABSTRACT: In the description of the electrostatic interaction between a solute treated at coupled cluster (CC) level of theory and a solvent modeled as a continuum dielectric, the solvent response depends on various contributions: the choice of the reference wave function, the correlation density, and the orbital relaxation. In previous work, we examined the first two factors with the coupled cluster singles and doubles (CCSD) method and its variant Brueckner doubles (BD) method. The CC wave function was combined with the polarizable continuum model (PCM) of solvation in an integrated and efficient method able to describe energy and molecular properties through analytic energy gradients. Additionally, we investigated some approximations, and proposed new ones, that reduce the computational cost to nearly that of gas phase CC while keeping most of the complete model description. In this work, we study the contribution of the orbital relaxation and compare it to the other effects. Such contribution is introduced with a self-consistent macroiteration procedure, where the reaction field is updated with a refined density. The results presented here show that the effect of the orbital relaxation is small for CCSD, while for BD the integrated and self-consistent approaches are equivalent. Thus, these results further confirm that the integrated CCSD-PCM and BD-PCM methods, especially with their respective approximations, are an efficient approach to perform high-level electronic structure calculations in solution.

1. INTRODUCTION

The interest in introducing the solvent effect in coupled cluster (CC)^{1,2} calculations is evident by considering the increasing number of publications on the subject, see for example refs 3–17.

We investigated the ability of the polarizable continuum model (PCM)^{18–20} of solvation to provide such effect in a reliable way, while maintaining the computational burden close to that of gas phase CC. We originally followed the approach in ref 11 for coupled cluster singles and doubles (CCSD) called perturbation theory energy and density (PTED) scheme for historical reasons^{21,22} and its approximation called PTE. In CCSD-PCM-PTED, the solvent reaction field is separated in a contribution from the reference wave function [usually the Hartree–Fock (HF) wave function] and a correlation contribution, which nonlinearly depends on the reduced one particle density matrix (1PDM). Such dependence from the 1PDM makes the PTED scheme computationally much more expensive than gas phase CCSD, since it requires the evaluation of the Λ vector.^{23,24} In the PTE scheme, the correlation reaction field is neglected. This reduces the computational cost at the expense of the accuracy of the method.^{13,16,17} However, by considering the part of the 1PDM that does not depend on Λ , i.e., the singles CC amplitudes, 50–80% of the PTE-PTED difference can be recovered while keeping the same computational cost of the PTE scheme. This is called PTES scheme, where the S stands for singles.¹⁷ All these schemes work in the frozen orbital framework, where the reference wave function is decoupled from the correlation calculation, consistently with the most widely used CC methods for isolated molecules. Thus, also the solvent reference and correlation reaction fields are decoupled, and no orbital relaxation is introduced.

One way to take into account such relaxation and to couple the two parts of the reaction field is to use a variant of CCSD called Brueckner doubles (BD).^{25,26} The BD reference function is built such that the amplitudes for the single excitations are all zero. Starting from a guess reference (HF or other), the orbitals are rotated until the above condition is satisfied. BD thus introduces an orbital relaxation that is coupled with the CC expansion (which includes now only double excitations). The numerical tests on the BD-PCM method,¹⁶ for the PTED and PTE schemes, showed that the latter is a much better approximation of the complete model for BD than for CCSD while keeping the computational cost comparable to the gas phase case. Additionally, the larger insensitivity of BD with respect to the choice of the initial reference function (compared to CCSD) makes this method more appealing for cases where the HF wave function may be unstable also in solution.

Another way to introduce the orbital relaxation in a CC-PCM calculation is through an external iteration procedure. This is an old idea, related to the initial implementation of PCM,¹⁹ and has been recently revived by Impropa et al. in the context of time-dependent density functional theory.^{27,28} In short, the algorithm:

- (1) computes an initial CC total density;
- (2) computes the reaction field corresponding to this density;
- (3) checks for convergence of the reaction field;
- (4) if not converged, introduces the reaction field in the reference calculation and computes an updated density;
- (5) returns to point 2.

In this procedure, the relaxed CC density is used, including the orbital response. This is a self-consistent method and can be

Received: September 21, 2011

Published: November 08, 2011

applied to any method for which it is possible to compute the 1PDM,²⁹ provided that the appropriate definition of the free energy is used. Indeed, in this work it is applied to CCSD and BD, and the results are compared to the PTED approach described above. The external iteration procedure is also a PTED approach, since the reaction field depends in the correlation density. However, in order to distinguish it from the frozen orbital approach, we will use “SC” for “self-consistent” to indicate the external iteration procedure and will keep “PTED” for the frozen orbital scheme. PCM-SC is not a computationally efficient procedure but is simple to implement and provides useful benchmarking results.

The goal of this work is to compare the PTED and SC schemes, for CCSD and BD, in order to evaluate the effect of the orbital relaxation on the solvent reaction field. Additionally, the approximated PTE and PTES schemes are tested, in search for the best compromise between accuracy and computational cost.

The numerical tests show that the reference reaction field is the major component of the solvent response, as it should be expected. The correlation contribution, however, can be significant and should not be neglected. The orbital relaxation, on the other hand, gives a very small contribution to the reaction field compared to that of correlation. Furthermore, the latter can be very well approximated by using the PTE scheme for BD and the PTES scheme for CCSD, drastically reducing the computational effort. These results indicate that CCSD-PCM-PTES and BD-PCM-PTE are very promising approaches to perform high-level calculations in solution at a cost comparable to the corresponding gas phase methods.

The paper is organized as follows: In Section 2, the free energy expressions for the PTED and SC schemes are presented and compared. The numerical applications are reported in Section 3, while an overall discussion and concluding remarks are the content of Section 4.

2. THEORY

In this section, we compare the free energy³⁰ expressions for the PCM-PTED and PCM-SC schemes when combined with the CCSD and BD methods. For a review of PCM, the interested reader is referred to refs 18–22, and 31–36. for the combination of PCM and CCSD/BD, to refs 11, 13, 16, and 17 and for the PCM-SC scheme to refs 27 and 28.

As outlined in Section 1, the solvent reaction field (i.e., the apparent surface charges) is defined differently in the PTED and SC schemes. In the former, the reaction field is divided into two contributions: one that depends on the reference function and the nuclei, $\bar{\mathbf{Q}}^0$, and one that depends on the CC reduced 1PDM, $\bar{\mathbf{Q}}_N$.^{11,13,16,17} The total reaction field is the sum of the two contributions: $\bar{\mathbf{Q}}^{\text{PTED}} = \bar{\mathbf{Q}}^0 + \bar{\mathbf{Q}}_N$. For CCSD, no orbital response is included in the total reaction field, since the reference wave function is decoupled from the coupled cluster expansion. On the other hand, for BD, the “correlation” reaction field $\bar{\mathbf{Q}}_N$ includes the orbital response, since this is included in the 1PDM through the solution of the Λ equations.²⁶ The interaction of the “reference” reaction field $\bar{\mathbf{Q}}^0$ with the solute wave function is included in the Fock operator, while the interaction of $\bar{\mathbf{Q}}_N$ introduces explicit terms in the CC equations (for BD, also $\bar{\mathbf{Q}}^0$ introduces some explicit term).¹⁶ In the SC scheme, the total reaction field, $\bar{\mathbf{Q}}^{\text{SC}}$, is computed all together from the nuclei and the total 1PDM, which includes the orbital response.^{27,28} This is obtained, for CCSD, by solving the appropriate coupled

perturbed HF (CPHF) equations.^{23,24,35} The interaction of $\bar{\mathbf{Q}}^{\text{SC}}$ with the solute wave function is entirely introduced in the Fock operator, and no explicit PCM term appears in the CC equations.

Keeping in mind the difference between $\bar{\mathbf{Q}}^{\text{PTED}}$ and $\bar{\mathbf{Q}}^{\text{SC}}$ described above, we can now introduce and discuss the expressions for the free energy in the two schemes. For PTED, the free energy is

$$G^{\text{PTED}} = \langle \Phi^0 | (1 + \Lambda) e^{-T} H_0 e^T | \Phi^0 \rangle + \frac{1}{2} \langle \Phi^0 | (1 + \Lambda) e^{-T} \mathbf{V} e^T | \Phi^0 \rangle \cdot \bar{\mathbf{Q}}^{\text{PTED}} \quad (1)$$

where Φ^0 is the reference function, H_0 is the molecular Hamiltonian, T is the excitation operator, Λ is the \mathbf{Z} vector from gradient theory,^{23,24} and \mathbf{V} is the electrostatic potential on the cavity surface. By using the normal product form of an operator (indicated with the N subscript),² the reference contribution to the free energy can be separated from the correlation contribution, so that eq 1 can be rewritten as

$$\begin{aligned} G^{\text{PTED}} &= G_0^{\text{PTED}} + \langle \Phi^0 | (1 + \Lambda) e^{-T} H_{0N} e^T | \Phi^0 \rangle \\ &+ \langle \Phi^0 | (1 + \Lambda) e^{-T} \mathbf{V}_N e^T | \Phi^0 \rangle \cdot \bar{\mathbf{Q}}^0 \\ &+ \frac{1}{2} \langle \Phi^0 | (1 + \Lambda) e^{-T} \mathbf{V}_N e^T | \Phi^0 \rangle \cdot \bar{\mathbf{Q}}_N \\ &= G_0^{\text{PTED}} + \langle \Phi^0 | (1 + \Lambda) e^{-T} H_{0N}^{\text{PTED}} e^T | \Phi^0 \rangle \\ &+ \frac{1}{2} \langle \Phi^0 | (1 + \Lambda) e^{-T} \mathbf{V}_N e^T | \Phi^0 \rangle \cdot \bar{\mathbf{Q}}_N \quad (2) \end{aligned}$$

where

$$G_0^{\text{PTED}} = \langle \Phi^0 | H_0 | \Phi^0 \rangle + \frac{1}{2} \langle \Phi^0 | \mathbf{V} | \Phi^0 \rangle \cdot \bar{\mathbf{Q}}^0 \quad (3)$$

is the reference free energy and

$$H_{0N}^{\text{PTED}} = H_{0N} + \mathbf{V}_N \cdot \bar{\mathbf{Q}}^0 \quad (4)$$

At convergence, eq 2 reduces to^{11,13,16,17}

$$\begin{aligned} G^{\text{PTED}} &= G_0^{\text{PTED}} + \langle \Phi^0 | e^{-T} H_{0N}^{\text{PTED}} e^T | \Phi^0 \rangle \\ &+ \langle \Phi^0 | e^{-T} \mathbf{V}_N e^T | \Phi^0 \rangle \cdot \bar{\mathbf{Q}}_N \\ &- \frac{1}{2} \langle \Phi^0 | (1 + \Lambda) e^{-T} \mathbf{V}_N e^T | \Phi^0 \rangle \cdot \bar{\mathbf{Q}}_N \\ &= G_0^{\text{PTED}} + \Delta E^{\text{PTED}} - \frac{1}{2} \bar{\mathbf{V}}_N \cdot \bar{\mathbf{Q}}_N \quad (5) \end{aligned}$$

where

$$\begin{aligned} \Delta E^{\text{PTED}} &= \langle \Phi^0 | e^{-T} H_{0N}^{\text{PTED}} e^T | \Phi^0 \rangle \\ &+ \langle \Phi^0 | e^{-T} \mathbf{V}_N e^T | \Phi^0 \rangle \cdot \bar{\mathbf{Q}}_N \quad (6) \end{aligned}$$

Equations 1–6 apply to both CCSD and BD, although for the latter, the last term in eq 6 vanishes since the singles amplitudes are zero.¹⁶

For the PCM-SC scheme, the free energy is

$$\begin{aligned} G^{\text{SC}} &= \langle \Phi^0 | (1 + \Lambda) e^{-T} H_0 e^T | \Phi^0 \rangle \\ &+ \frac{1}{2} \langle \Phi^0 | (1 + \Lambda) e^{-T} \mathbf{V} e^T | \Phi^0 \rangle \cdot \bar{\mathbf{Q}}^{\text{SC}} \quad (7) \end{aligned}$$

Note that, for CCSD, the expression in eq 7 is different from that used in refs 27 and 28, since the orbital response is only explicitly

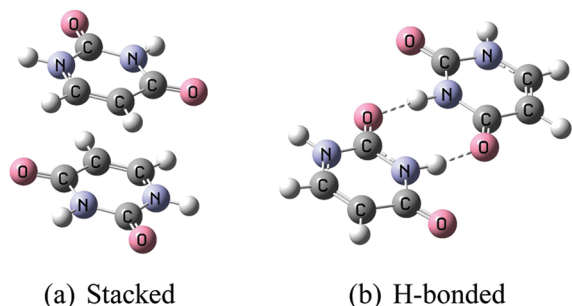


Figure 1. Structure of the uracil dimers.

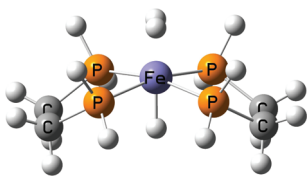


Figure 2. Structure of the $\text{H}_2[\text{FeH}(\text{PP})_2]^+$ complex.

included in the apparent charges and not in the potential. Using the definition in eq 7, G^{SC} is consistent with the equations for the CC amplitudes. On the other hand, for BD, eq 7 includes the orbital response through the Λ vector and is equivalent to refs 27 and 28. Separating the reference contribution from the correlation one, eq 7 becomes

$$G^{\text{SC}} = G_0^{\text{SC}} + \langle \Phi^0 | (1 + \Lambda) e^{-T} H_{0N} e^T | \Phi^0 \rangle + \frac{1}{2} \langle \Phi^0 | (1 + \Lambda) e^{-T} \mathbf{V}_N e^T | \Phi^0 \rangle \cdot \bar{\mathbf{Q}}^{\text{SC}} \quad (8)$$

where

$$G_0^{\text{SC}} = \langle \Phi^0 | H_0 | \Phi^0 \rangle + \frac{1}{2} \langle \Phi^0 | \mathbf{V} | \Phi^0 \rangle \cdot \bar{\mathbf{Q}}^{\text{SC}} \quad (9)$$

Using the CC amplitudes equation in the presence of the reaction field:

$$\langle \Phi | e^{-T} H_{0N}^{\text{SC}} e^T | \Phi^0 \rangle = 0 \quad (10)$$

where Φ is an excited determinant and

$$H_{0N}^{\text{SC}} = H_{0N} + \mathbf{V}_N \cdot \bar{\mathbf{Q}}^{\text{SC}} \quad (11)$$

eq 8 can be simplified to

$$\begin{aligned} G^{\text{SC}} &= G_0^{\text{SC}} + \langle \Phi^0 | e^{-T} H_{0N}^{\text{SC}} e^T | \Phi^0 \rangle \\ &\quad - \frac{1}{2} \langle \Phi^0 | (1 + \Lambda) e^{-T} \mathbf{V}_N e^T | \Phi^0 \rangle \cdot \bar{\mathbf{Q}}^{\text{SC}} \\ &= G_0^{\text{SC}} + \Delta E^{\text{SC}} - \frac{1}{2} \bar{\mathbf{V}}_N \cdot \bar{\mathbf{Q}}^{\text{SC}} \end{aligned} \quad (12)$$

where

$$\Delta E^{\text{SC}} = \langle \Phi^0 | e^{-T} H_{0N}^{\text{SC}} e^T | \Phi^0 \rangle \quad (13)$$

As mentioned in Section 1, G^{SC} is computed with a self-consistent procedure: several PCM macroiterations are performed where the charges $\bar{\mathbf{Q}}^{\text{SC}}$ are calculated with an updated density and reentered in the reference and CC equations. The iterative procedure is stopped when convergence on the reaction field is achieved.

Table 1. Solvation Free Energy (kcal/mol) for CCSD^a

	exptl ^b	PTE ^c	PTES	PTED ^c	SC
Pyridine					
Cyc	-4.30	-1.76	-1.78	-1.78	-1.79
Dce	-5.53	-4.04	-4.08	-4.11	-4.11
H ₂ O	-4.70	-4.73	-4.77 ^d	-4.80	-4.80
Aniline					
Cyc	-5.52	-1.92	-1.94	-1.94	-1.95
Dce	-7.39	-4.51	-4.54	-4.57	-4.57
H ₂ O	-5.49	-5.32	-5.35 ^d	-5.38	-5.38
Phenol					
Cyc	-5.57	-2.14	-2.16	-2.16	-2.17
Dce	-7.48	-4.94	-4.97	-4.99	-4.99
H ₂ O	-6.62	-5.79	-5.82 ^d	-5.85	-5.85
<i>p</i> -Bromophenol					
Cyc	-7.14	-2.39	-2.42	-2.42	-2.43
Dce	-9.10	-5.40	-5.44	-5.46	-5.46
H ₂ O	-7.13	-6.29	-6.33 ^d	-6.36	-6.35
Chlorobenzene					
Cyc	-5.10	-0.95	-0.97	-0.98	-0.98
Dce		-2.15	-2.18	-2.20	-2.20
H ₂ O	-1.12	-2.50	-2.53 ^d	-2.57	-2.57

^a Cyc: cyclohexane, Dce: dichloroethane, H₂O: water. ^b Ref 39. ^c Ref 13. ^d Ref 17.

Equations 5 and 12 look very similar. For the CCSD method, the difference is that the SC scheme includes the orbital response in the reaction field, and the reference and CC equations are coupled by $\bar{\mathbf{Q}}^{\text{SC}}$. In the PTED scheme, the reference and the post-HF part are decoupled, which is consistent with the gas phase CC method. For the BD method, the PTED and SC schemes are equivalent since in BD the reference and the CC expansion are already coupled, and the orbital response is taken into account in both schemes.

3. RESULTS

A variety of systems is used to compare the PTED and SC schemes. The solvation free energy, defined as the difference between the free energy in solution and the gas phase energy, is computed for five organic molecules: pyridine, aniline, phenol, *p*-bromophenol, and chlorobenzene, in three solvents of increasing polarity: cyclohexane, dichloroethane, and water. Two dimers of uracil in stacked and hydrogen bonded (H-bonded) conformation, see Figure 1, are used to compare the interaction energy in gas phase and water and the solvation free energy. Finally, a dihydrogen complex of Fe⁺ in tetrahydrofuran (THF), see Figure 2, is analyzed by comparing the solvation free energy at different H–H bond lengths. The PCM cavity is built from interlocking spheres centered on each nucleus and using the universal force field (UFF) radii.³⁷ The symmetric version of integral equation formalism PCM (IEF-PCM)³⁶ is used, and Scalmani and Frisch's continuous surface charge (CSC) scheme³⁵ is employed. All the calculations are performed with a development version of the Gaussian suite of programs.³⁸

In this section, results for two approximations of the PTED model are also presented. One approximation is called PTE

Table 2. Solvation Free Energy (kcal/mol) for BD^a

	exptl ^b	PTE ^c	PTED ^c	SC
Pyridine				
Cyc	-4.30	-1.79	-1.80	-1.79
Dce	-5.53	-4.10	-4.11	-4.10
H ₂ O	-4.70	-4.79	-4.80	-4.79
Aniline				
Cyc	-5.52	-1.94	-1.94	-1.94
Dce	-7.39	-4.54	-4.55	-4.55
H ₂ O	-5.49	-5.35	-5.36	-5.36
Phenol				
Cyc	-5.57	-2.17	-2.17	-2.17
Dce	-7.48	-4.98	-4.99	-4.99
H ₂ O	-6.62	-5.83	-5.84	-5.84
<i>p</i> -Bromophenol				
Cyc	-7.14	-2.42	-2.42	-2.42
Dce	-9.10	-5.44	-5.45	-5.45
H ₂ O	-7.13	-6.33	-6.34	-6.34
Chlorobenzene				
Cyc	-5.10	-0.97	-0.98	-0.98
Dce		-2.19	-2.19	-2.19
H ₂ O	-1.12	-2.55	-2.55	-2.55

^a Cyc: cyclohexane, Dce: dichloroethane, H₂O: water. ^b Ref 39. ^c Ref 16.

scheme, for CCSD and BD,^{11,13,16} where the correlation reaction field contribution is neglected. The other is the PTES scheme (where S stands for singles), for CCSD,¹⁷ where only part of the CCSD reduced 1PDM is used to compute the correlation reaction field. Both these approximated schemes do not require the calculation of the complete CC reduced 1PDM, thus decoupling the *T* and Λ equations. This strongly reduces the computational cost, making the PCM-PTE and -PTES methods similar to gas phase CC.

The results for the solvation free energies of the five molecules are reported in Table 1 for CCSD and Table 2 for BD. The tables also report the experimental data.³⁹ The geometries are taken from ref 13 for CCSD and ref 16 for BD. The 6-31+G(d,p) basis set is used. The calculations in Tables 1 and 2 basically show no difference between PTED and SC, as all of the results are within 0.01 kcal/mol from each other. For CCSD, the difference between PTE and PTED, which is a measure of the correlation reaction field effect, although already small (at most 0.07 kcal/mol), is larger than the orbital response effect. Additionally, such difference can be greatly reduced with the PTES scheme ($\geq 50\%$). For BD, PTE is able to recover most of the correlation solvent effect through the coupling of the reference function and the CC expansion, without introducing the explicit correlation reaction field. The comparison with experiment also shows that nonelectrostatic effects, neglected in the calculations, are far more important for low-polar solvents than other effects. As shown in refs 13 and 16, such contributions can be recovered by adding the proper correction³⁹ to the PCM electrostatic free energy. Therefore, for this first set of molecules, the orbital response is the smallest of all the contributions to the free energy.

Tables 3–6 collect the data computed for the uracil dimers in stacked and H-bonded conformations (see Figure 1). The geometry for the stacked conformation is taken from ref 40,

Table 3. CCSD Interaction and Solvation Free Energies (kcal/mol) for the Stacked Uracil Dimer^a

	interaction energy				
	gas	PTE	PTES	PTED	SC
6-31G(d)	-2.43	0.30	0.39	0.42	0.39
6-31+G(d,p)	-3.67	-0.41	-0.33	-0.31	-0.33
solvation free energy					
	PTE	PTES	PTED	SC	
Dimer					
6-31G(d)	-22.63	-23.01	-23.10	-23.10	-22.97
6-31+G(d,p)	-25.36	-25.64	-25.71	-25.71	-25.63
Monomer 1					
6-31G(d)	-12.68	-12.92	-12.97	-12.97	-12.90
6-31+G(d,p)	-14.31	-14.49	-14.54	-14.54	-14.48
Monomer 2					
6-31G(d)	-12.68	-12.92	-12.97	-12.97	-12.90
6-31+G(d,p)	-14.31	-14.49	-14.54	-14.54	-14.48

^a The interaction energies include a BSSE correction.

Table 4. BD Interaction and Solvation Free Energies (kcal/mol) for the Stacked Uracil Dimer^a

	interaction energy			
	gas	PTE	PTED	SC
6-31G(d)	-2.44	0.43	0.43	0.43
6-31+G(d,p)	-3.64	-0.29	-0.28	-0.29
solvation free energy				
	PTE	PTED	SC	
Dimer				
6-31G(d)	-23.02	-23.05	-23.05	-23.05
6-31+G(d,p)	-25.67	-25.70	-25.70	-25.69
Monomer 1				
6-31G(d)	-12.94	-12.96	-12.96	-12.96
6-31+G(d,p)	-14.51	-14.52	-14.52	-14.52
Monomer 2				
6-31G(d)	-12.94	-12.96	-12.96	-12.96
6-31+G(d,p)	-14.51	-14.52	-14.52	-14.52

^a The interaction energies include a BSSE correction.

and the monomers are 3.3 Å apart. The geometry for the H-bonded dimer is taken from ref 41 and reoptimized at the B3LYP level^{42,43} with the aug-cc-pVDZ basis set⁴⁴ imposing *C_s* symmetry. The calculations are performed with the 6-31G(d) and 6-31+G(d,p) basis sets. The tables report the interaction energies, in gas and in solution, and the solvation free energies in kcal/mol. The interaction energies include a correction for the basis set superposition error (BSSE) through the counterpoise method.^{45,46} The same BSSE correction is used for the gas phase and PCM calculations, since there is not a well-defined procedure

Table 5. CCSD Interaction and Solvation Free Energies (kcal/mol) for the H-Bonded Uracil Dimer^a

	interaction energy				
	gas	PTE	PTES	PTED	SC
6-31G(d)	-10.32	-5.53	-5.48	-5.47	-5.49
6-31+G(d,p)	-10.53	-5.12	-5.08	-5.07	-5.08
solvation free energy					
	PTE	PTES	PTED	SC	
Dimer					
6-31G(d)	-20.90	-21.32	-21.43	-21.29	
6-31+G(d,p)	-23.61	-23.93	-24.02	-23.92	
Monomer 1					
6-31G(d)	-12.85	-13.09	-13.15	-13.08	
6-31+G(d,p)	-14.52	-14.71	-14.76	-14.70	
Monomer 2					
6-31G(d)	-12.83	-13.07	-13.13	-13.05	
6-31+G(d,p)	-14.50	-14.67	-14.72	-14.67	

^aThe interaction energies include a BSSE correction.

to compute this quantity with continuum solvation models. These basis sets are not large enough for quantitative results, as shown by a comparison of the gas phase interaction energies with the best estimates in the literature:^{29,41,40} -12.4 and -7.5 kcal/mol for the H-bonded and the stacked dimers, respectively. There is a difference of 2–4 kcal/mol with the values in this work (see Tables 3–6). However, here we are only interested in the comparison between the various PCM approaches. The solvent competes with the pair interaction, so that the interaction energy is reduced by ~3 kcal/mol for the stacked conformation and ~5 kcal/mol for the H-bonded one. For CCSD (Tables 3 and 5), the difference in the interaction energies between the PTED and SC schemes is about 0.01–0.03 kcal/mol with both basis sets. This is again smaller than the PTE-PTED difference, which is already small (0.04–0.12 kcal/mol). The solvation free energy difference between the PTE and PTED schemes is larger, around 0.4–0.5 kcal/mol for the dimer and 0.2–0.3 kcal/mol for the monomers. On the other hand, the effect of the orbital relaxation produces a difference of only 0.08–0.14 kcal/mol for the dimer and 0.06–0.07 kcal/mol for the monomers. As in the previous example, the PTES scheme strongly reduced the PTE-PTED difference to 0.01–0.02 kcal/mol for the interaction energy, 0.07–0.11 kcal/mol for the solvation free energy of the dimers, and 0.05–0.06 kcal/mol for the monomers, while keeping the computational cost of the PTE scheme. For BD, there is no difference between the PTED and the SC approaches, as in the previous example. Furthermore, PTE is a very good approximation, since the difference with PTED is at most 0.01 kcal/mol for the interaction energy and 0.03 kcal/mol for the solvation energy, see Tables 4 and 6.

For the $\text{H}_2[\text{FeH}(\text{PP})_2]^+$ complex, the geometry is taken from ref 16. The calculations are performed with the Stuttgart/Dresden pseudopotential⁴⁷ for Fe, the 6-31G(d,p) basis set for the dihydrogen and the hydrogen in trans position, and the 6-31G(d) basis for the rest of the atoms. An extra sphere is included along the Fe–H₂ bond to avoid “solvent pockets” at longer H–H distances (the sphere diameter being equal to the Fe–H₂

Table 6. BD Interaction and Solvation Free Energies (kcal/mol) for the H-Bonded Uracil Dimer^a

	interaction energy			
	Gas	PTE	PTED	SC
6-31G(d)	-10.34	-5.46	-5.45	-5.45
6-31+G(d,p)	-10.55	-5.08	-5.08	-5.08
solvation free energy				
	PTE	PTED	SC	
Dimer				
6-31G(d)	-21.33	-21.36	-21.36	
6-31+G(d,p)	-23.96	-23.98	-23.98	
Monomer 1				
6-31G(d)	-13.12	-13.13	-13.13	
6-31+G(d,p)	-14.73	-14.74	-14.74	
Monomer 2				
6-31G(d)	-13.10	-13.11	-13.11	
6-31+G(d,p)	-14.70	-14.71	-14.71	

^aThe interaction energies include a BSSE correction.

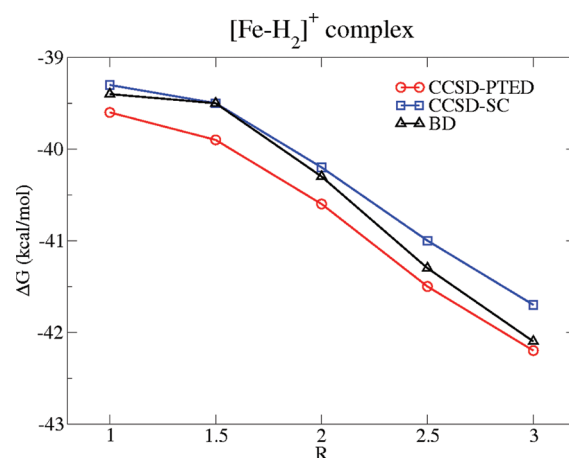


Figure 3. Solvation free energy (kcal/mol) for the $\text{H}_2[\text{FeH}(\text{PP})_2]^+$ complex at various H–H bond distances. R is a scaling factor from the equilibrium bond distance.

distance: 1.6219 Å). Figure 3 shows the change in solvation free energy when scaling the equilibrium H–H distance (0.8105 Å) by a factor R . For BD, only one curve is reported since the results with the SC and PTED schemes are equivalent. The difference between the two schemes for CCSD is at most 0.4–0.5 kcal/mol. This is ~3 times smaller than the correlation effect (1.1–1.4 kcal/mol),¹⁶ and in the opposite direction. In our previous work, we showed that the approximated CCSD-PCM-PTES and BD-PCM-PTE schemes recover most, if not all, of this effect (compare Figure 6 in ref 16 and Figure 2 in ref 17).

4. DISCUSSION AND CONCLUSIONS

The test examples in Section 3 show that, for CCSD, the orbital relaxation is the smallest among all the contributions to

the solvent reaction field. The most important one is the contribution from the reference wave function, as expected. However, the correlation contribution can be significant, as for the uracil dimers, where it reaches 0.5 kcal/mol, and even more the Fe complex, where it is of the order of 1.1–1.4 kcal/mol.¹⁶ In the latter case, the orbital relaxation contribution may amount to a third of the correlation one.

Although this effect may seem significant, it comes at the prize of a computationally less efficient, and somewhat less elegant formulation. Indeed, in the self-consistent macroiteration procedure of the CCSD-PCM-SC scheme,^{27,28} the coupling of the reference wave function and the post-HF part are only provided by the reaction field, and not through a variational optimization of the CC free energy functional. Additionally, the inclusion of the orbital relaxation effect makes the formulation of free energy derivatives complicated (it requires the derivative of the CC density), and expensive, thus limiting its applicability for molecular properties in solution. The advantage of this strategy, however, is the rather straightforward implementation, and the possibility to use it with any method for which the IPDM is available.

The PTED scheme, on the other hand, is based on the decoupling of the reference function and the correlation contributions to the reaction field. This is consistent with CC for isolated molecules. It also allows for an efficient formulation of the free energy analytic gradients, and therefore, it can be efficiently used to compute molecular properties in solution. As mentioned above, this approach neglects the contribution of the orbital relaxation to the reaction field, but this effect is relatively small.

For the BD variant of CCSD, the SC and PTED schemes are equivalent, since the orbital response is taken into account in both. This is an advantage of this method over CCSD, especially when the HF wave function may be unstable.¹⁶ However, BD is computationally more expensive than CCSD, and the latter can be used in the majority of cases.

Other effects that can be more important than the orbital response are the nonelectrostatic contributions to the free energy, when low-polar solvents are used, see Tables 1 and 2. These can be accounted for in a variety of ways, and we showed an example in refs 13 and 16. In the same papers, we also mentioned that explicit solute–solvent interactions, like strong H-bonds, can play an important role that cannot be accurately reproduced by continuum solvation models.

Computational efficiency can be obtained with careful approximations of the PTED model, for CCSD and BD, through the PTES and PTE schemes, respectively. Both these approximated schemes recover most of the correlation reaction field contribution (50–80% with CCSD-PCM-PTES and ~100% with BD-PCM-PTE), while keeping the computational cost comparable with the correspondent gas phase methods.

In conclusion, although the orbital response can provide a nonnegligible contribution to the free energy, this remains rather small compared to the other contributions (reference and correlation). Additionally, the frozen orbitals CCSD-PCM-PTED scheme is consistent with gas phase CCSD, and analytical free energy gradients can be efficiently formulated with small modification of the gas phase formulas and computer code. Therefore, the latter can be considered more appropriate for CC calculation in solution with PCM. In light of all of these results, the best compromise between accuracy and computational effort is achieved with the PTES approximation for CCSD¹⁷ and the PTE approximation for BD.¹⁶

AUTHOR INFORMATION

Corresponding Author

*E-mail: marco@gaussian.com.

REFERENCES

- (1) Bartlett, R. J.; Musial, M. *Rev. Mod. Phys.* **2007**, *79*, 291–352.
- (2) Shavitt, I.; Bartlett, R. J. *Many-Body Methods in Chemistry and Physics*; Cambridge University Press: Cambridge, U.K., 2009.
- (3) Christiansen, O.; Mikkelsen, K. V. *J. Chem. Phys.* **1999**, *110*, 1365–1375.
- (4) Christiansen, O.; Mikkelsen, K. V. *J. Chem. Phys.* **1999**, *110*, 8348–8360.
- (5) Kongsted, J.; Osted, A.; Pedersen, T. B.; Mikkelsen, K. V.; Christiansen, O. *J. Phys. Chem. A* **2004**, *108*, 8624–8632.
- (6) Osted, A.; Kongsted, J.; Mikkelsen, K. V.; Christiansen, O. *J. Phys. Chem. A* **2004**, *108*, 8646–8658.
- (7) Aidas, K.; Kongsted, J.; Osted, A.; Mikkelsen, K. V.; Christiansen, O. *J. Phys. Chem. A* **2005**, *109*, 8001–8010.
- (8) Snegov, K.; Schwabe, T.; Kongsted, J.; Christiansen, O. *J. Chem. Phys.* **2011**, *134*, 104108.
- (9) Slipchenko, L. V. *J. Phys. Chem. A* **2010**, *114*, 8824–8830.
- (10) Kosenkov, D.; Slipchenko, L. V. *J. Phys. Chem. A* **2011**, *115*, 392–401.
- (11) Cammi, R. *J. Chem. Phys.* **2009**, *131*, 164104.
- (12) Caricato, M.; Mennucci, B.; Scalmani, G.; Trucks, G. W.; Frisch, M. J. *J. Chem. Phys.* **2010**, *132*, 084102.
- (13) Caricato, M.; Scalmani, G.; Trucks, G. W.; Frisch, M. J. *J. Phys. Chem. Lett.* **2010**, *1*, 2369–2373.
- (14) Cammi, R.; Fukuda, R.; Ehara, M.; Nakatsuji, H. *J. Chem. Phys.* **2010**, *133*, 024104.
- (15) Fukuda, R.; Ehara, M.; Nakatsuji, H.; Cammi, R. *J. Chem. Phys.* **2011**, *134*, 104109.
- (16) Caricato, M.; Scalmani, G.; Frisch, M. J. *J. Chem. Phys.* **2011**, *134*, 244113.
- (17) Caricato, M. *J. Chem. Phys.* **2011**, *135*, 074113.
- (18) Miertus, S.; Scrocco, E.; Tomasi, J. *J. Chem. Phys.* **1981**, *55*, 117–129.
- (19) Tomasi, J.; Persico, M. *Chem. Rev.* **1994**, *94*, 2027.
- (20) Tomasi, J.; Mennucci, B.; Cammi, R. *Chem. Rev.* **2005**, *105*, 2999–3093.
- (21) Olivares del Valle, F.; Tomasi, J. *J. Chem. Phys.* **1991**, *150*, 139–150.
- (22) Aguilar, M.; Olivares del Valle, F.; Tomasi, J. *J. Chem. Phys.* **1991**, *150*, 151–161.
- (23) Salter, E. A.; Trucks, G. W.; Bartlett, R. J. *J. Chem. Phys.* **1989**, *90*, 1752–1766.
- (24) Gauss, J.; Stanton, J. F.; Bartlett, R. J. *J. Chem. Phys.* **1991**, *95*, 2623–2638.
- (25) Handy, N. C.; Pople, J. A.; Head-Gordon, M.; Raghavachari, K.; Trucks, G. W. *J. Chem. Phys. Lett.* **1989**, *164*, 185–192.
- (26) Kobayashi, R.; Handy, N. C.; Amos, R. D.; Trucks, G. W.; Frisch, M. J.; Pople, J. A. *J. Chem. Phys.* **1991**, *95*, 6723–6733.
- (27) Impropa, R.; Barone, V.; Scalmani, G.; Frisch, M. J. *J. Chem. Phys.* **2006**, *125*, 054103.
- (28) Impropa, R.; Scalmani, G.; Frisch, M. J.; Barone, V. *J. Chem. Phys.* **2007**, *127*, 074504.
- (29) Lipparini, F.; Scalmani, G.; Mennucci, B. *J. Phys. Chem. Chem. Phys.* **2009**, *11*, 11617–11623.
- (30) Ho, J.; Klamt, A.; Coote, M. L. *J. Phys. Chem. A* **2010**, *114*, 13442–13444.
- (31) Cancès, E.; Mennucci, B.; Tomasi, J. *J. Chem. Phys.* **1997**, *107*, 3032–3041.
- (32) Mennucci, B.; Cancès, E.; Tomasi, J. *J. Phys. Chem. B* **1997**, *101*, 10506–10517.
- (33) Klamt, A.; Schüürmann, G. *J. Chem. Soc., Perkin Trans. 2* **1993**, 799–805.

- (34) Cossi, M.; Rega, N.; Scalmani, G.; Barone, V. *J. Comput. Chem.* **2003**, *24*, 669–681.
- (35) Scalmani, G.; Frisch, M. J. *J. Chem. Phys.* **2010**, *132*, 114110.
- (36) Lipparini, F.; Scalmani, G.; Mennucci, B.; Cancès, E.; Caricato, M.; Frisch, M. J. *J. Chem. Phys.* **2010**, *133*, 014106.
- (37) Rappe, A. K.; Casewit, C. J.; Colwell, K. S.; Goddard, W. A.; Skiff, W. M. *J. Am. Chem. Soc.* **1992**, *114*, 10024–10035.
- (38) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A. Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Keith, T.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Parandekar, P. V.; Mayhall, N. J.; Daniels, A. D.; Farkas, O.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. *Gaussian Development Version*, revision H.09+; Gaussian, Inc.: Wallingford, CT, 2010.
- (39) Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. B* **2009**, *113*, 6378–6396.
- (40) Šponer, J.; Jurečka, P.; Marchan, I.; Luque, F. J.; Orozco, M.; Hobza, P. *Chem.—Eur. J.* **2006**, *12*, 2854–2865.
- (41) Šponer, J.; Jurečka, P.; Hobza, P. *J. Am. Chem. Soc.* **2004**, *126*, 10142–10151.
- (42) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648–5652.
- (43) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. *J. Chem. Phys.* **1994**, *98*, 11623–11627.
- (44) Dunning, T. H. *J. Chem. Phys.* **1989**, *90*, 1007–1023.
- (45) Boys, S. F.; Bernardi, F. *Mol. Phys.* **1970**, *19*, 553–566.
- (46) Simon, S.; Duran, M.; Dannenberg, J. J. *J. Chem. Phys.* **1996**, *105*, 11024–11031.
- (47) Dolg, M.; Wedig, U.; Stoll, H.; Preuss, H. *J. Chem. Phys.* **1987**, *86*, 866–872.

Incorporating Linear Synchronous Transit Interpolation into the Growing String Method: Algorithm and Applications

Andrew Behn,[†] Paul M. Zimmerman,[‡] Alexis T. Bell,[†] and Martin Head-Gordon^{*,†}

[†]Department of Chemical and Biomolecular Engineering and [‡]Department of Chemistry, University of California, Berkeley, California 94720-1462, United States

ABSTRACT: The growing string method is a powerful tool in the systematic study of chemical reactions with theoretical methods which allows for the rapid identification of transition states connecting known reactant and product structures. However, the efficiency of this method is heavily influenced by the choice of interpolation scheme when adding new nodes to the string during optimization. In particular, the use of Cartesian coordinates with cubic spline interpolation often produces guess structures which are far from the final reaction path and require many optimization steps (and thus many energy and gradient calculations) to yield a reasonable final structure. In this paper, we present a new method for interpolating and reparameterizing nodes within the growing string method using the linear synchronous transit method of Halgren and Lipscomb. When applied to the alanine dipeptide rearrangement and a simplified cationic alkyl ring condensation reaction, a significant speedup in terms of computational cost is achieved (30–50%).

INTRODUCTION

One of the key contributions of theoretical chemistry to the systematic study of chemical reactions is the ability to accurately predict kinetic rate constants. These kinetic rate constants are typically calculated with transition state theory, which requires knowledge of the transition state structure. While locating stable minima on the potential energy surface (PES) is considered relatively easy in theoretical chemistry, the automated location of first order transition states remains a challenge.

The principle method for obtaining exact first order transition states connecting known reactant and product configurations is to first generate a rough guess of the structure and then refine this structure to the exact answer through surface walking.^{1–4} The algorithms for surface walking are similar to the algorithms which locate PES minima. Because there are many more transition states than minima on a typical PES, this guess must be very close (within the basin of attraction) to the proper transition state in order to properly converge. Once the transition state has been refined, it must be confirmed by integrating the reaction path downhill to the reactant and product configurations.^{5,6}

Several algorithms for finding transition state guesses from known reactant and product configurations have been developed,^{7–33} including the nudged elastic band method (NEB),^{5–8} the string method (SM),^{11–16} and the growing string method (GSM).^{17–22} In each of these “chain-of-states” methods,²³ the minimum energy pathway is located by iteratively optimizing a discretized representation of the pathway. Each of the nodes in the chain-of-states is a full molecular structure at some intermediate stage of the transition between the reactant and product. Optimization steps are taken by moving each image downhill on the PES, perpendicular to the direction of the reaction path. Additionally, the nodes in the chain are kept equally spaced through the optimization process, either through an additional spacing force or by explicit reparameterization. This ensures that this node-based description of the pathway does not contain large gaps, where the PES may be left unsampled.

When using *ab initio* surfaces in each of these methods, the overall cost of generating a suitable guess of the transition state can be stated in terms of the overall number of QM nuclear gradient calculations performed. All other calculations needed to perform these methods can be considered negligible. Higher order derivatives of the QM energy, such as the nuclear Hessian, would provide faster convergence but are typically expensive in *ab initio* calculations.

The most commonly encountered chain-of-states method is the nudged elastic band method,^{7,8} which finds an approximate reaction path by optimizing a series of images connected to each other through a set of springs with contrived hooke constants. The optimization step direction for each node, v_{NEB}^i , is comprised of two components, as shown in eq 1.

$$\hat{v}_{\text{NEB}}^i = \frac{-g_i^\perp + f_i^\parallel}{|-g_i^\perp + f_i^\parallel|} \quad (1)$$

The first term is the perpendicular force, used to minimize the energy of each node, and is given by eq 2.

$$-g_i^\perp = -(I - \hat{t}_i \hat{t}_i^T) g_i \quad (2)$$

The tangent direction, \hat{t}_i , is typically found by normal finite difference, but other schemes have been proposed for improved performance.^{9,10} The second term, f_i^\parallel , expanded in eq 3, is a force along the reaction path which arises from the springs that connect each node in the chain to its neighbors. Here, k is the spring constant, and R denotes the coordinates of a molecular structure.

$$f_i^\parallel = \hat{t}_i \hat{t}_i^T [(R_{i+1} - R_i) - (R_i - R_{i-1})] k \quad (3)$$

This component is added to ensure that the images remain equally spaced during the optimization process.

Received: September 18, 2011

Published: November 16, 2011

A slightly newer method, similar in concept to the nudged elastic band, is the string method.^{11–16} The reaction path is again represented by a series of molecular images, but optimization is broken into two separate steps: evolution and reparameterization. In the evolution step, the molecular images are moved in the direction of the negative perpendicular gradient, similarly to the first term in eq 1. The tangent direction is determined by creating a cubic spline through each Cartesian coordinate of the string of images. The reparameterization step is performed by reinterpolating the molecular images along this set of cubic splines to achieve the desired parametrization density. This avoids the need to decide an arbitrary spring constant, as in the NEB.

The growing string method^{17–22} is a modification of the original string method that aims to reduce overall computational cost by “growing” the set of nodes from the reactant and product configurations inward toward the transition state. In principle, this avoids performing gradient calculations on excessively rugged parts of the PES that are far from the final reaction pathway. Initially, the string consists of only the reactant and product configurations, with one node being added to each side during the first reparameterization. The string is then evolved and reparameterized in the normal fashion until the specified convergence criteria for node addition are achieved by the innermost nodes on the reactant and product sides. New nodes are added accordingly until the string is fully populated. The string then continues to optimize until convergence is reached. Several schemes have been developed to accelerate practical use by improving optimization, and using cost-saving dual basis techniques.^{18–20}

The nudged elastic band and string method both require initial pathways from which the optimizations are launched. This initial pathway has a tremendous impact on convergence and must be chosen carefully.¹⁴ Cartesian interpolation may work for some simple reactions; however, it is not always an appropriate choice. Reactions such as the HNC to HCN isomerization are poorly described by Cartesian interpolation despite having only a handful of internal degrees of freedom. Such simple reactions, as well as more complex reactions with many atoms, require an interpolation scheme with the chemical intuition built-in. GSM does not require a full initial guess pathway but does require a methodology for reparameterizing and interpolating new nodes.

To alleviate this shortcoming of GSM specifically (and more generally the SM and NEB), we propose using linear synchronous transit (LST) interpolation^{34,35} for node interpolation and reparameterization. LST is a method for interpolating between two fixed molecular geometries that seeks to preserve internuclear distances within the molecule as it morphed from one to the other. In doing so, the usual drawbacks of Cartesian interpolation in which chemical bonds are overly compressed or stretched are avoided. A simple example of the contrast between LST and Cartesian interpolation is demonstrated in Figure 1 with the HCN to HNC isomerization. In this paper, we demonstrate the advantages of LST interpolation in GSM by direct comparison to Cartesian interpolation.

MODIFIED IMPLEMENTATION OF THE GROWING STRING METHOD

I. Evolution Step. To demonstrate the use of LST versus Cartesian interpolation within the GSM, a modified version of the algorithm was developed. The evolution step is performed by moving each current node in the string, i , along the negative

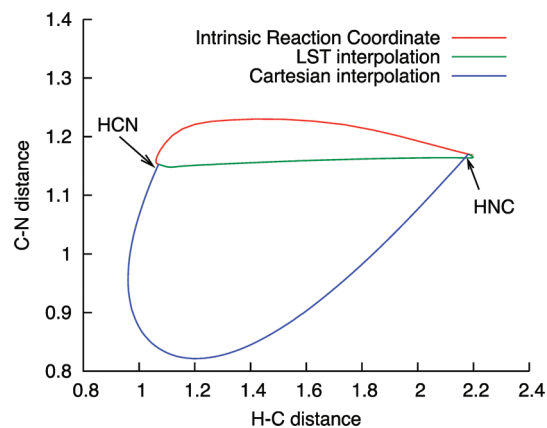


Figure 1. Contrast of Cartesian and LST interpolation between reactant and product configurations of the HCN to HNC reaction. Note that the Cartesian interpolated pathway is far from the minimum energy path, while the LST interpolated pathway is rather close. In addition, the LST method’s tendency to preserve internuclear distances is clearly shown.

perpendicular gradient, as shown in eq 4. The tangent direction is determined for each node with either LST interpolation or by computing the cubic splines over the set of Cartesian coordinates of the string.

$$\hat{v}_i = \frac{-(I - t_i t_i^T)g_i}{|(I - t_i t_i^T)g_i|} = \frac{-g_i^\perp}{|g_i^\perp|} \quad (4)$$

The length of each node’s evolution step is computed by dividing the magnitude of the node’s perpendicular gradient by a common scaling factor, γ , as shown in eq 5. This produces a damped steepest descent step and has the effect of generating a large step when there is a large perpendicular gradient far from convergence and a small step when the node is near convergence.

$$d_i = \frac{|g_i^\perp|}{\gamma} \quad (5)$$

The overall step, Δx_i , as a combination of eqs 4 and 5 is presented in eq 6.

$$\Delta x_i = \hat{v}_i d_i = \frac{-g_i^\perp}{\gamma} \quad (6)$$

This step differs from the original GSM,¹⁷ which seeks to minimize the string by taking several small trial steps in the downhill direction, fitting the observed energy profile to a quadratic function, and moving to the estimated minimum.

II. Reparameterization Step. After each evolution step, the string is reparameterized to achieve a uniform node density along the arclength of the reaction path. If nodes are numbered starting with the reactant node as $i = 1$ and the product node as $i = N$, where N is the number of nodes in the fully populated string, the desired parametrization is given by eq 7. N_R and N_P are the number of nodes on the reactant and product sides respectively, and s_{tot} is the current total arclength from reactant to product. The exact computation of s_{tot} is done with the appropriate interpolation scheme (discussed below).

$$s_i = s_{\text{tot}} \left(\frac{i-1}{N-1} \right) \text{ for } i \leq N_R \text{ and } N - N_P < i \leq N \quad (7)$$

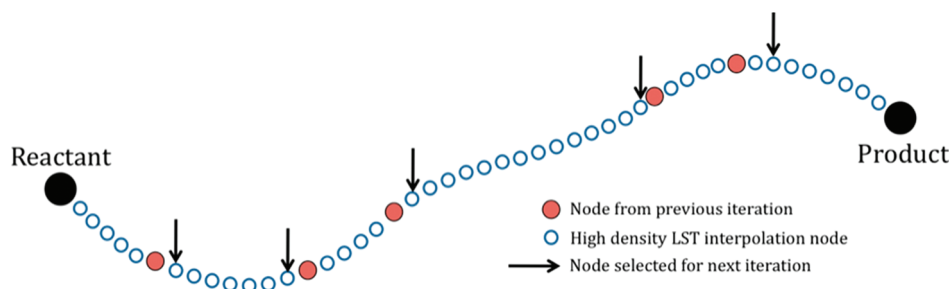


Figure 2. Cartoon of the strategy for using LST interpolation within the GSM. The spaces between the previous iteration nodes are filled with a discrete set of LST interpolated images. The set of interpolated images which returns the appropriate parameterization density is then returned for the next iteration.

The algorithm is started with two nodes on each side of the string—the fixed reactant and product structures and one variable node on each side. Once the magnitude of the perpendicular gradient for the innermost node for a side has fallen below the specified tolerance for node addition, a new node is added to a side by incrementing N_R or N_P during the reparameterization step. This has the effect of growing a new node on the appropriate side until the string is fully populated. Once the full string has been grown, convergence may be considered. The objective function for locating the minimum energy path (MEP) is the sum of the perpendicular gradient magnitudes for each node, as indicated by eq 8. If the end points are assumed to be stable minima on the PES, they may be neglected in the sum since they contribute nothing.

$$F = \sum_{i=2}^{N-1} |g_i^\perp| \quad (8)$$

INTERPOLATION METHOD

If Cartesian coordinates with cubic splines are used for reparameterization, the procedure is straightforward. A cubic spline is determined for each Cartesian coordinate using the structures of the nodes and their positions along the string in terms of arclength, and the appropriate nodes are interpolated.

The use of linear synchronous transit reparameterization is slightly more complicated. It is based on the use of LST interpolation between two fixed molecular images, as given by the resultant structure in the minimization of eq 9.

$$G = \sum_{a>b}^{\text{atoms}} \frac{(r_{ab}^i - r_{ab}^c)^2}{(r_{ab}^i)^4} + 10^{-6} \sum_{a=1}^{\text{atoms}} \sum_{j=x,y,z} (w_{a,j}^i - w_{a,j}^c)^2 \quad (9)$$

The r variables denote internuclear distances, while the w variables denote pure Cartesian coordinates. The i and c superscripts denote “interpolated” versus “computed” values respectively. The interpolated values are determined by mixing the values of the fixed molecular structures, while the computed values are derived from the interpolated structure being optimized to minimize G . We are careful to stress that the “computed” internuclear distances, r_{ab}^c , are derived from the coordinates provided by the Cartesian “computed” structure, w^c . Thus, there is only one full set of Cartesian coordinates being manipulated. The numerator of the first term of eq 9 serves to preserve internuclear distances from being overly stretched or compressed during interpolation, while the denominator weights

this effect in favor of shorter internuclear distances (i.e., bonding interactions). The second term of eq 9 provides a small force to align the interpolated molecule with the fixed end structures.

Equation 9 fails to adequately show that the “interpolated” values must be computed by choosing a mixing ratio, f , of the two fixed structures. This is shown in eq 10, where the superscripts 1 and 2 denote the fixed end point structures.

$$\begin{aligned} r_{ab}^i &= r_{ab}^1 + f(r_{ab}^2 - r_{ab}^1) \\ w_{a,j}^i &= w_{a,j}^1 + f(w_{a,j}^2 - w_{a,j}^1) \end{aligned} \quad (10)$$

From these equations, it becomes obvious that G is really $G = G(f)$, and the choice of f (between 0 and 1) will determine how close the interpolated image is to the fixed end points. A value of $f = 0$ will produce an interpolated structure identical to structure 1, while $f = 1$ will reproduce structure 2. If all values of f between 0 and 1 are sampled and the LST equation minimized at each value, it yields a continuous description of the deformation from structure 1 to structure 2, with an integrated arclength of $s_{\text{LST}}(f)$. There exists a monotonically increasing but nonlinear mapping between f and $s_{\text{LST}}(f)$ such that it is impossible to know *a priori* which value of f to use to return a desired value of s_{LST} . To avoid this problem, a high-density series of LST interpolations must be performed between each neighboring set of nodes in the evolving string.

The general strategy for using LST interpolation for reparameterization in GSM is shown in Figure 2. First, a high-density set of LST interpolations is performed between each neighboring pair of nodes. From this, the normalized arclength position of each interpolated node along the growing string is computed. Finally, the nodes which yield the appropriate node spacing, as given by eq 7, are selected from the high-density string and taken as the reparameterized string. From this same high density LST interpolated string, the tangent vector at each selected node is computed and stored.

COMPUTATIONAL DETAILS

The examples detailed below demonstrate the use of Cartesian and LST interpolation in our modified implementation of the growing string method interfaced with Q-Chem 3.2.³⁶ For each example, a string of 11 nodes was grown from the reactant and product structures and optimized until a specified objective function was achieved. At the beginning of the GSM execution, the reactant and product structures were aligned to be in maximum coincidence in non mass-weighted Cartesian coordinates.³⁷ This step is essential to ensure that the rotational and translational degrees of freedom between the two structures

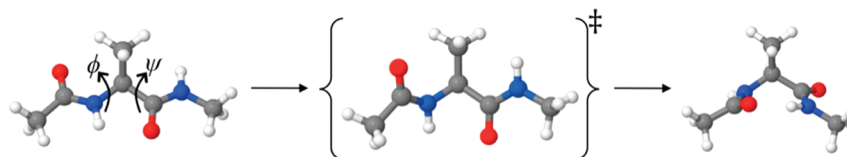


Figure 3. The reactant, transition state, and product configurations for the alanine–dipeptide rearrangement reaction.

Table 1. Dihedral Angles (in degrees) of the Alanine Dipeptide Isomerization

	ϕ	ψ
C_5	−161.5	167.1
TS	113.7	−141.9
C_{7AX}	72.5	−59.5

do not significantly affect the interpolation. It also ensures that the computed arclength between nodes does not include appreciable noninternal motion.

The scaling factor used to generate the step length of each node during the evolution step was $\gamma = 5.0$ hartree/Å². Reparameterization was performed after each evolution step, with new nodes added to the string (during the reparameterization steps) if the magnitude of the perpendicular gradient at a frontier node fell below 0.1 hartree/Å. For reparameterization and node addition with LST, 200 images were included in each high-density interpolation string running from reactant to product. The structures of this high-density string were optimized with Newton–Raphson minimization of eq 9 to a tolerance of $|\nabla G| < 0.001$ when computed in units of Ångstroms.

After the string was fully optimized, the nodes of the string were used as the starting point of a surface walking transition state optimization calculation in Q-Chem. This algorithm, which operates in delocalized internal coordinates, seeks to maximize the energy along the eigenvector of the lowest Hessian eigenmode and minimize the energy along all other eigenmodes. To aid in these calculations, an exact Hessian was calculated at the outset of the search and updated via the Powell/Murtagh–Sargent scheme.^{38,39}

Once a first-order saddle point was isolated from these optimizations, a high-quality MEP was integrated downhill from the transition state, via the Schlegel–Gonzalez MEP following algorithm⁶ in non-mass-weighted Cartesian coordinates, to ensure that the transition state connected the reactant and product structures initially fed to the growing string method. It is possible that multiple transition states may be found if each node is used to launch a calculation. For elementary reaction steps, only the highest energy node for an adequately converged string should result in a meaningful transition state. For nonelementary reaction steps, legitimate transition states may be found for each elementary reaction. Both of these possibilities are explored in the test cases presented below.

■ ALANINE DIPEPTIDE REARRANGEMENT

A common test problem for the benchmarking of MEP and transition state finding algorithms is the rearrangement of alanine dipeptide from the C_5 isomer to the C_{7AX} isomer. The minimum energy pathway involves the concerted rotation of the two dihedral angles, ϕ and ψ , shown in Figure 3. The relevant values of the dihedral angles for the reactant, TS, and product in the gas

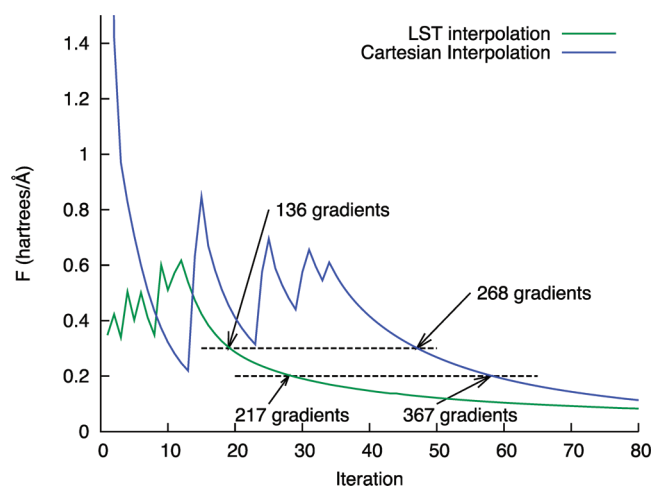


Figure 4. Objective function, F , vs iteration for the alanine dipeptide rearrangement reaction. The spikes in the initial portion of each curve indicate the addition of new nodes during the growth phase.

phase at the B3LYP/6-31G level of theory given in the work of Perczel et al.⁴⁰ are shown in Table 1. This is the same level of theory used in the present example.

Figure 4 shows the value of the objective function F (from eq 8) as a function of iteration. These curves can each be broken into two regions: growth and refinement. The initial growth phase, during which new nodes are still being added to the string, results in the spikes seen initially in Figure 4. Since the number of nodes in each iteration is not constant in this phase, the number of QM gradients necessary for each iteration also varies. The subsequent refinement phase begins once the string has been fully grown and is marked by the monotonically decreasing value of F during which the string settles into the reaction pathway. For alanine dipeptide rearrangement with Cartesian interpolation, the growth phase is completed after the 33rd iteration, corresponding to 133 QM gradient calculations. With LST interpolation, growth is completed after the 11th iteration and 55 QM gradient calculations (see Figure 5 for a comparison of the intermediate energy profiles for the LST and Cartesian GSM). The growth phase is much faster with LST due to the superior quality of the new interpolated nodes. This demonstrated in Figure 6, which shows the energy as a function of iteration for the first node added to the reactant side of the string. The LST interpolated node begins at a much lower energy and achieves the threshold for the next node addition more quickly.

Table 2 lists the number of QM gradient calculations necessary to fully grow the string and achieve the desired level of convergence. For a convergence criterion of 0.3 hartree/Å, LST interpolation reduces the number of QM gradients required by 49%, effectively doubling the speed of GSM. The string energy profiles for each interpolation method at a convergence of $F = 0.3$

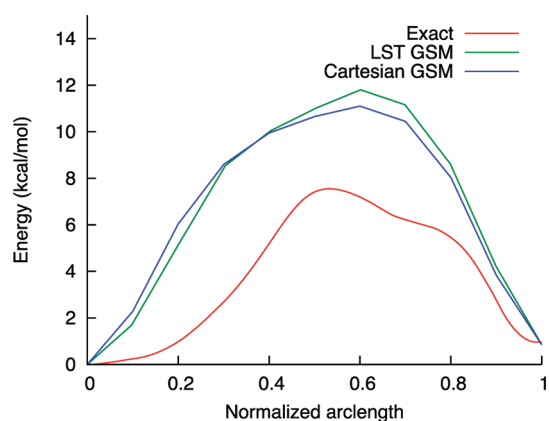


Figure 5. Comparison of the intermediate energy profiles for the LST and Cartesian GSM and the exact MEP energy profile for the alanine dipeptide rearrangement. The GSM energy profiles are snapshots taken when the objective function, F , reached 0.3 hartree/Å.

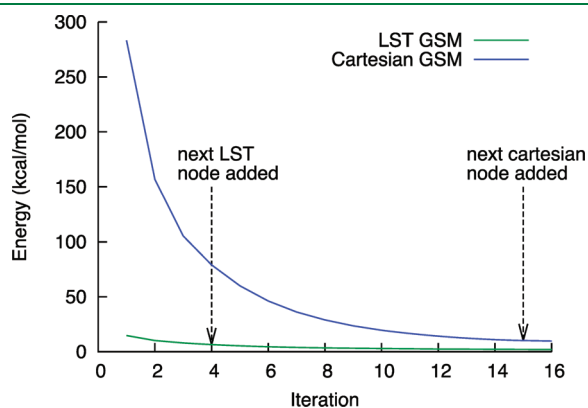


Figure 6. Energy of the first interior node from the reactant side versus iteration for the alanine dipeptide rearrangement. Note that the LST interpolated node begins at a much lower energy than the Cartesian interpolated node and achieves the threshold for node addition much sooner.

Table 2. Computational Cost in QM Gradients for the Alanine Dipeptide Rearrangement, with Speedup for LST versus Cartesian Interpolation

conv. criteria	method	QM gradients	speedup
0.3 hartree/Å	Cartesian	268	49%
	LST	136	
0.2 hartree/Å	Cartesian	367	41%
	LST	217	

hartree/Å shown in Figure 6 indicate that the overall quality of the string with Cartesian and LST interpolation is approximately the same. For a more tightly converged reaction coordinate at $F = 0.2$ hartree/Å, the speedup is similar at 41%. For each of the four strings resulting from LST and Cartesian interpolation methods in GSM at these two convergence criteria, the highest energy node yields the proper transition state using the standard Q-Chem surface-walking algorithm detailed in the Computational Details section.

RING CONDENSATION REACTION

The second benchmarking case is the cationic ring condensation of 2-(but-3-enyl)oxiranium to 4-hydroxycyclohexan-1-ylum. This reaction is inspired by the much more complicated (and enzymatically catalyzed) reaction of 2,3-oxidosqualene to produce lanosterol during cholesterol synthesis.⁴¹ Even though this reaction is simplified, it possesses four transition states on the path from linear reactant to final product at the gas-phase HF/STO-3g level of theory. Each of these transition states and the stable intermediates are depicted in Figure 7. TS1, TS2, and TS4 are each internal rotations of the molecule with small barriers between 2.0 and 3.0 kcal/mol. The remaining transition state, TS3, involves the rearrangement of bond orders and possesses a much higher barrier of 23.2 kcal/mol.

Figure 8 shows the convergence rate of GSM with both Cartesian and LST interpolation for this reaction. With Cartesian interpolation, the growth phase is completed after 36 iterations, which corresponds to 186 QM gradient calculations. With LST interpolation, the growth phase requires only 16 iterations,

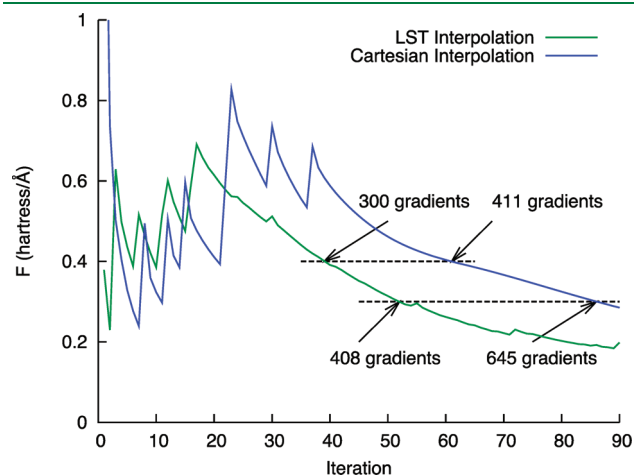


Figure 8. Objective function, F , versus iteration for the cationic ring condensation reaction. The spikes in the initial portion of each curve indicate the addition of nodes during the growth phase.

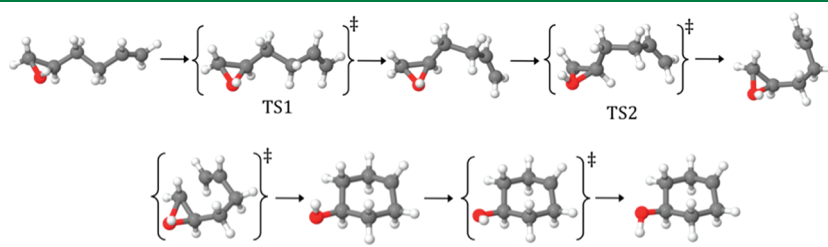
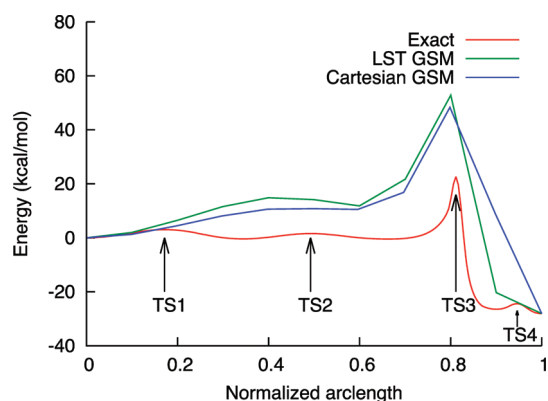


Figure 7. Reactant, product, stable intermediates, and transition states on the pathway between the linear and ring structures.

Table 3. Computational Costs, Speedups, and Success in Identifying Various TS Structures for the Cationic Ring Condensation Reaction

conv. criteria	method	QM gradients	speedup	TS1	TS2	TS3	TS4
0.4 hartree/Å	Cartesian	411	27%	Y	Y	Y	N
	LST	300		Y	Y	Y	N
0.3 hartree/Å	Cartesian	645	37%	Y	Y	Y	N
	LST	408		Y	Y	Y	Y

**Figure 9.** Comparison of the intermediate energy profiles for the LST and Cartesian GSM and the exact MEP energy profile for the ring condensation reaction. The GSM energy profiles are snapshots taken when the objective function, F , reached 0.3 hartree/Å.

corresponding to 84 QM gradient calculations. As denoted in Table 3, a 27% reduction in the number of gradients is observed for convergence to $F = 0.4$ hartree/Å, and a 37% reduction is observed for convergence to $F = 0.3$ hartree/Å. Figure 9 shows the exact energy profile for this reaction, as well as the energy profiles for both GSM executions once the objective function reached 0.3 hartree/Å.

In each of the four cases noted in Table 3, launching TS optimization calculations in Q-Chem³⁶ from the converged strings' nodes resulted in several first order transition states. In all cases, TS1, TS2, and TS3 from Figure 7 were successfully recovered. However, TS4 (which involves a subtle hydroxyl group rotation) was only observed for the more tightly converged ($F = 0.3$ hartree/Å) LST interpolated string.

CONCLUSIONS

The growing string method is a powerful tool in the study of chemical reactions from an *ab initio* perspective because it allows for the rapid identification of transition states, from which approximate kinetic rate constants may be computed with transition state theory. However, the interpolation method by which the string is reparameterized and new nodes are added during the growth phase has a large impact on the rate of convergence and thus the quality of results. In particular, choosing an interpolation scheme which compresses or expands chemical bonds arbitrarily can necessitate a large number of QM calculations in order to properly find the minimum energy path and transition state.

Our results indicate that using the linear synchronous transit method developed initially by Halgren and Lipscomb³⁴ can be a powerful addition to the traditional growing string method. This

interpolation method is an improvement over Cartesian interpolation because it preserves bond lengths and performs rotational rearrangements seamlessly. The guessed pathways are thus closer to the final result and require less computational effort to optimize.

When applied to the isomerization of alanine dipeptide, GSM with LST interpolation requires roughly half of the computational effort as GSM with Cartesian interpolation. In the condensation of 2-(but-3-enyl)oxiranium to 4-hydroxycyclohexan-1-ylum, computational cost is reduced by roughly one-third when LST interpolation is used. In this latter reaction, which contains multiple transition states between the reactant and product, the LST version of the GSM proves superior by properly identifying every transition state (major and minor) where the Cartesian version misses at least one.

AUTHOR INFORMATION

Corresponding Author

*E-mail: mhg@cchem.berkeley.edu.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENT

This work was supported by the Methane Conversion Cooperative funded by BP. Calculations were performed on a cluster provided by the UC Berkeley College of Chemistry through grant NSF CHE-0840505.

REFERENCES

- (1) Cerjan, C. J.; Miller, W. H. *J. Chem. Phys.* **1981**, *112*, 2129.
- (2) Banerjee, A.; Adams, N.; Simons, J.; Shepard, R. J. *Phys. Chem.* **1985**, *89*, 52.
- (3) Hranchian, H. P.; Schlegel, H. B. *J. Comput. Chem.* **2003**, *24*, 1514.
- (4) Schlegel, H. B. *J. Comput. Chem.* **2003**, *24*, 1514.
- (5) Fukui, K. *J. Phys. Chem.* **1970**, *74*, 4161.
- (6) Gonzalez, C.; Schlegel, H. B. *J. Chem. Phys.* **1989**, *90*, 2154.
- (7) Mills, G.; Jonsson, H. *Phys. Rev. Lett.* **1994**, *72*, 1124.
- (8) Henkelman, G.; Jonsson, H. *J. Chem. Phys.* **2000**, *113*, 9978.
- (9) Henkelman, G.; Uberuaga, B. P.; Jonsson, H. *J. Chem. Phys.* **2000**, *113*, 9901.
- (10) Trygubenko, S. A.; Wales, D. J. *J. Chem. Phys.* **2004**, *120*, 2083.
- (11) E, W.; Ren, W.; Vanden-Eijnden, E. *Phys. Rev. B.* **2002**, *66*, 052301.
- (12) E, W.; Ren, W.; Vanden-Eijnden, E. *J. Phys. Chem. B.* **2005**, *109*, 6688.
- (13) E, W.; Ren, W.; Vanden-Eijnden, E. *J. Chem. Phys.* **2007**, *126*, 164103.
- (14) Sheppard, D.; Terrell, R.; Henkelman, G. *J. Chem. Phys.* **2008**, *128*, 134106.
- (15) Burger, S. K.; Yang, W. *J. Chem. Phys.* **2006**, *124*, 054109.
- (16) Burger, S. K.; Yang, W. *J. Chem. Phys.* **2007**, *127*, 164107.
- (17) Peters, B.; Heyden, A.; Bell, A. T. *J. Chem. Phys.* **2004**, *120*, 7877.
- (18) Goodrow, A.; Bell, A. T.; Head-Gordon, M. *J. Chem. Phys.* **2008**, *129*, 174109.
- (19) Goodrow, A.; Bell, A. T.; Head-Gordon, M. *J. Chem. Phys.* **2009**, *130*, 244108.
- (20) Goodrow, A.; Bell, A. T.; Head-Gordon, M. *Chem. Phys. Lett.* **2010**, *484*, 393.
- (21) Quapp, W. *J. Comput. Chem.* **2007**, *28*, 1834.
- (22) Quapp, W. *J. Theor. Comput. Chem.* **2009**, *8*, 101.
- (23) Elber, R.; Karplus, M. *Chem. Phys. Lett.* **1987**, *139*, 375.

- (24) Ayala, P. Y.; Schlegel, H. B. *J. Chem. Phys.* **1997**, *107*, 375.
- (25) del Campo, J. M.; Koster, A. M. *J. Chem. Phys.* **2008**, *129*, 024107.
- (26) Ghasemi, S. A.; Goedecker, S. J. *J. Chem. Phys.* **2011**, *135*, 014108.
- (27) Burger, S. K.; Ayers, P. W. *J. Chem. Phys.* **2010**, *132*, 234110.
- (28) Maeda, S.; Morokuma, K. *J. Chem. Phys.* **2010**, *132*, 241102.
- (29) Dey, B. K.; Ayers, P. W. *Mol. Phys.* **2006**, *104*, 541.
- (30) Aguilar-Mogas, A.; Gimenez, X.; Bofill, J. M. *J. Chem. Phys.* **2008**, *128*, 104102.
- (31) Aguilar-Mogas, A.; Gimenez, X.; Bofill, J. M. *J. Comput. Chem.* **2010**, *31*, 2510.
- (32) Klimes, J.; Bowler, D. R.; Michaelides, A. *J. Phys: Condens. Mater.* **2010**, *22*, 074203.
- (33) Koslover, E. F.; Wales, D. J. *J. Chem. Phys.* **2007**, *127*, 134102.
- (34) Halgren, T. A.; Lipscomb, W. N. *Chem. Phys. Lett.* **1977**, *49*, 225.
- (35) Peng, C.; Schlegel, H. B. *Israel J. Chem.* **1993**, *33*, 449.
- (36) Shao, Y.; et al. *Phys. Chem. Chem. Phys.* **2006**, *8*, 3172.
- (37) Rhee, Y. M. *J. Chem. Phys.* **2000**, *113*, 6021.
- (38) Powell, M. J. D. *Math. Prog.* **1971**, *1*, 26.
- (39) Murtagh, B. A.; Sargent, R. W. H. *Comput. J.* **1972**, *13*, 185.
- (40) Perczel, A.; Farkas, O.; Jakli, I.; Topol, I. A.; Csizmadia, I. G. *J. Comput. Chem.* **2002**, *24*, 1026.
- (41) Wendt, K. U.; Schulz, G. E.; Corey, E. J.; Liu, D. R. *Angew. Chem., Int. Ed.* **2000**, *39*, 2812.

An Automated Force Field Topology Builder (ATB) and Repository: Version 1.0

Alpeshkumar K. Malde,^{*,†} Le Zuo,[†] Matthew Breeze,[†] Martin Stroet,[†] David Poger,[†] Pramod C. Nair,[†] Chris Oostenbrink,^{‡,§} and Alan E. Mark^{*,†,||}

[†]School of Chemistry and Molecular Biosciences, University of Queensland, St. Lucia, Australia

[‡]Leiden/Amsterdam Center for Drug Research, Division of Molecular Toxicology, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

[§]Institute of Molecular Modeling and Simulation, University of Natural Resources and Life Sciences, Vienna, Austria

^{||}Institute for Molecular Bioscience, University of Queensland, St. Lucia, QLD 4072, Australia

S Supporting Information

ABSTRACT: The Automated force field Topology Builder (ATB, <http://compbio.biosci.uq.edu.au/atb>) is a Web-accessible server that can provide topologies and parameters for a wide range of molecules appropriate for use in molecular simulations, computational drug design, and X-ray refinement. The ATB has three primary functions: (1) to act as a repository for molecules that have been parametrized as part of the GROMOS family of force fields, (2) to act as a repository for pre-equilibrated systems for use as starting configurations in molecular dynamics simulations (solvent mixtures, lipid systems pre-equilibrated to adopt a specific phase, etc.), and (3) to generate force field descriptions of novel molecules compatible with the GROMOS family of force fields in a variety of formats (GROMOS, GROMACS, and CNS). Force field descriptions of novel molecules are derived using a multistep process in which results from quantum mechanical (QM) calculations are combined with a knowledge-based approach to ensure compatibility (as far as possible) with a specific parameter set of the GROMOS force field. The ATB has several unique features: (1) It requires that the user stipulate the protonation and tautomeric states of the molecule. (2) The symmetry of the molecule is analyzed to ensure that equivalent atoms are assigned identical parameters. (3) Charge groups are assigned automatically. (4) Where the assignment of a given parameter is ambiguous, a range of possible alternatives is provided. The ATB also provides several validation tools to assist the user to assess the degree to which the topology generated may be appropriate for a given task. In addition to detailing the steps involved in generating a force field topology compatible with a specific GROMOS parameter set (GROMOS 53A6), the challenges involved in the automatic generation of force field parameters for atomic simulations in general are discussed.

INTRODUCTION

Computer simulations are widely used to gain insight into dynamic molecular processes at near atomic resolution. In particular, molecular dynamics (MD) methods can be used to model the time evolution of biomolecular systems (proteins, lipids, nucleic acids, and carbohydrates) in order to study dynamic processes such as protein folding or to sample a Boltzmann-weighted ensemble to estimate thermodynamic properties such as ligand binding free energies. The primary challenge in such simulations is to describe the properties of the system in terms of the interactions between atoms. Given the size and complexity of biomolecular systems together with the time scales that must be reached to model processes of interest, the use of classical mechanics in conjunction with a so-called *effective* force field is normally the method of choice. In such MD simulations, the interactions between the various particles in a given system are represented by an empirical energy function, parametrized to reproduce a range of structural, energetic, or thermodynamic properties of model compounds derived from experimental and/or high level quantum mechanical calculations.

A variety of such empirical force fields have been developed for use in biomolecular simulations. These include the GROMOS,^{1–5} AMBER,^{6,7} CHARMM,^{8,9} and OPLS^{10,11} force fields. Although the form of the potential energy function used in each of these

force fields is very similar, the parameters that are used to describe specific molecules can differ significantly. This in part reflects different parametrization philosophies. However, it also reflects the fact that the range of data that can be used during the parametrization is limited. As a result, parametrization is an underdetermined problem. Force field development is also further complicated by the fact that the parameters themselves are highly correlated. Common force fields such as GROMOS,^{1–5} AMBER,^{6,7} CHARMM,^{8,9} and OPLS^{10,11} provide a set of internally consistent parameters for a core set of molecules such as amino acids, nucleotides, simple sugars, common lipids, and common solvents. These are not general force fields per se but have instead been specifically parametrized to reproduce a given set of properties of this core set of molecules. As a consequence, each new heteromolecule such as a substrate, inhibitor, cofactor, or drug molecule must be parametrized individually. A number of more general force fields intended for the description of a wide range of small molecules have also been developed. Examples include MM2,¹² MM3,¹³ MM4,^{14,15} and MMFF94.^{16–18} Such force fields employ more complex potential energy functions expressed in terms of atomic properties and, as such, are

Received: March 22, 2011

Published: October 27, 2011

incompatible with most of the biomolecular force fields in common use.

The most basic approach that can be used to develop a force field description for a novel molecule compatible with one of the major biomolecular force fields is to manually assign parameters to groups of atoms based on their similarity to groups of atoms in molecules that have already been specifically parametrized for that force field. This is, however, tedious, time-consuming, and error-prone. Also, to be done appropriately, a detailed knowledge of the philosophy underlying a given force field is required. As a consequence, various tools have been developed to facilitate the generation of the force field descriptions for novel molecules for a range of force fields. A number of these generate simplified topologies and parameters that can be used to maintain a particular geometry of a ligand molecule for use in X-ray refinement. For example, the program eLBOW¹⁹ (electronic Ligand Builder and Optimization Workbench), a module of the PHENIX²⁰ suite of programs, generates a topology, parameters, and geometric restraints for ligand molecules based on the results of semiempirical quantum mechanical (AM1²¹) calculations. It uses a rule-based approach to identify all possible bonds, bond angles, dihedral angles, planar rings, and chiral centers in a molecule. Hydrogen atoms are added and semiempirical QM calculations performed to optimize the geometry of the molecule in a vacuum. The optimized geometry is then used to derive a set of purely geometric parameters for X-ray refinement in the CCP4²² monomer library format. Note that this format does not contain information regarding the force constants used to impose these geometric restraints. The program LIBCHECK,²³ distributed as part of the CCP4²² package, has a database of molecules that can be searched for known ligands based on element type and bonding patterns. In cases when no appropriate ligand is found, the parameters are generated using a rule-based approach using the geometry supplied by the user. The new molecule is then added to the database. LIBCHECK is also used by the refinement programs Refmac²⁴ and Coot.²⁵ XPLO2D, part of the X-PLOR and CNS^{26,27} packages, generates a highly simplified topology for use with the program CNS based on a set of coordinates supplied by the user. Again, a set of rules is used to identify atom type, bonds, bond angles, etc., which are combined with a small set of default values for the force constants to generate a set of geometric restraints. For example, 4184 kJ mol⁻¹ Å⁻² is used for all bond lengths, 2092 kJ mol⁻¹ rad⁻² for all bond angles, and 3140 kJ mol⁻¹ for all proper and improper dihedral angles of constrained systems and 0 kJ mol⁻¹ in all other cases. The HIC-Up²⁸ (Heterocompound Information Centre—Uppsala, <http://xray.bmc.uu.se/hicup/>) Web server provides parameter and topology files generated using XPLO2D for all heteromolecules found in the PDB and is updated once or twice a year. The program Hess2FF²⁹ can also be used to generate a topology and geometric constraints in CNS format. Hess2FF uses a Hessian (force constant matrix) of a molecule as input, which can be obtained from molecular mechanics, semiempirical or quantum mechanical calculations. The molecular descriptions generated by the tools listed above are, however, intended for use in X-ray refinement and are not suitable for molecular simulations. Many do not provide terms to model the nonbonded atoms (electrostatic and van der Waals interactions) or only provide parameters for heavy, non-hydrogen atoms. While these simplified descriptions are widely used, it is increasingly apparent that errors in the description of ligand molecules, often linked to the use of these very simplified representations, lead to errors and

uncertainties in structural databases such as the Protein Data Bank (PDB). This potentially has severe implications for structure-based drug design, interpretation of biochemical mechanisms, and validation of virtual screening.³⁰

A number of programs and Web servers have been developed to facilitate the generation of force field descriptions of novel molecules for use in MD simulations. MKTOP³¹ (<http://labmm.iq.ufrj.br/mktop/>) implements a rule-based approach to generate an OPLS-AA like force field in a GROMACS-readable format. A distance criterion of 0.3 Å is used to identify bonded partners. This information is used subsequently to assign atom types based on chemical environment and to derive other bonded parameters. The program does not assign partial charges to the atoms. Instead, these must be supplied by the user together with a set of coordinates for atoms in the ligand. Antechamber³² (<http://ambermd.org/antechamber/antechamber.html>) and YASARA AutoSMILES Server (<http://www.yasara.org/autosmilesserver.htm>) generate topologies compatible with the GAFF⁷ (Generalized AMBER force field) based on a set of coordinates supplied by the user. In the program Antechamber, first the atom types and atomic charges are assigned. If the user supplies a QM output file containing the electrostatic potential (ESP), then the partial atomic charges are assigned on the basis of the RESP³³ method; otherwise AM1-BCC³⁴ [Bond Charge Correction (BCC) applied to the AM1 atomic charges] charges are assigned. The bonded parameters are assigned on the basis of the GAFF parameters subsequently. In the YASARA server, an AM1 calculation is performed to optimize the geometry of the ligand. Initially, AM1-BCC parameters (including charges) are assigned to the molecule. However, if fragments within the molecule match fragments within a standard GAFF database, the charges are replaced by RESP charges, and the bonded parameters replaced by standard GAFF parameters. GENRTF³⁵ (<http://a.cmm.ki.si/genrtf/>) generates parameters and a topology compatible with the CHARMM force field using a rule-based approach. SwissParam (<http://swissparam.ch/>) generates a force field description for small organic molecules by combining bonded parameters extracted from the Merck Molecular Force Field (MMFF) and nonbonded terms from the CHARMM22 force field. Paramchem server (<https://www.paramchem.org>) also aims to generate parameters consistent with the CHARMM force field. PRODRG³⁶ (<http://davapc1.bioch.dundee.ac.uk/cgi-bin/prodrgr/prodrgr.cgi>) is a Web server which generates force field descriptions of ligand molecules based roughly on the GROMOS 43A1 force field. The molecular topologies generated by PRODRG are intended for use in X-ray refinement, docking, and MD simulations. Nevertheless, PRODRG has a number of serious limitations, including that the ligand protonation states are assigned automatically by PRODRG, the assignment of critical 1–4 exclusions is inappropriate in some cases, and atomic charges and charge groups are not assigned in a manner consistent with the GROMOS force field.³⁷ In fact, it is not possible to assign the protonation/tautomeric state and/or the overall charge of the molecule using many of the tools and Web servers described above. Although the YASARA AutoSMILES Server does consider the pH when adding hydrogen atoms to a given structure, the tautomeric state cannot be assigned.

Here, we describe the Automated force field Topology Builder (ATB, <http://compbio.biosci.uq.edu.au/atb>) and repository. The ATB is a Web-accessible server that can provide topologies and parameters for a wide range of molecules compatible with the GROMOS family of force fields. The topologies and parameters

provided are designed for use in molecular simulations, computational drug design, and X-ray refinement. The ATB has three primary functions: first, to act as a repository for molecules that have been parametrized as part of the GROMOS family of force fields; second, to act as a repository for pre-equilibrated systems for use as starting configurations in molecular dynamics simulations (solvent mixtures, lipid systems pre-equilibrated to adopt a specific phase, etc.); and third, to generate force field descriptions of novel molecules compatible with the GROMOS force field in a variety of formats. This is done in a multistep process in which results from a series of quantum mechanical (QM) calculations are combined with a knowledge-based approach to ensure compatibility with the GROMOS family of force fields. The ATB differs from other topology builders in that it requires the user to stipulate the protonation and tautomeric states of the molecule. The symmetry of the molecule is also analyzed to ensure equivalent atoms are assigned identical parameters irrespective of molecular geometry. Charge groups, which are used in the GROMOS force field to ensure that compensating groups of charges are always considered simultaneously, thus reducing artifacts in the calculation of the long-range electrostatic interactions, are assigned automatically. Importantly, in cases where the assignment of a given parameter is ambiguous, a range of possible alternatives is provided. Finally, the ATB provides several validation tools to assist the user to assess the degree to which the topology generated may be appropriate for a given task.

The remaining sections of the manuscript are organized as follows. First, the challenges associated with the automatic generation of molecular force fields are discussed briefly. Then, the basic structure of the ATB pipeline is presented. This is followed by a discussion on the limitations of the current version of the ATB along with the steps taken to validate the final topologies.

CHALLENGES IN THE AUTOMATIC GENERATION OF MOLECULAR FORCE FIELDS

The empirical potential energy functions used in biomolecular force fields such as GROMOS,^{1–5} AMBER,^{6,7} CHARMM,^{8,9} and OPLS^{10,11} are crude in the sense that they attempt to represent the potential energy surfaces of a wide range of molecules using a very limited set of parameters. Differences in chemical environment are encoded by assigning different sets of van der Waals interaction parameters for a given atom type depending on the neighboring atoms. Pair interactions are based on simple combination rules. Electrostatic interactions are modeled by assigning fixed partial charges to atoms. In addition, the parameters are often correlated. For example, the choice of van der Waals parameters is correlated with the partial charge, and dihedral terms are correlated with both the other bonded as well as nonbonded parameters. This leads to a number of challenges when attempting to generate a molecular force field description automatically. To select an appropriate atom type, one must first be able to identify the local chemical environment, determine if the atom is aromatic or aliphatic, if needed determine whether the atom may be a hydrogen-bond donor or acceptor, and assign a partial charge to an atom. In particular, assigning appropriate partial atomic charges is challenging. The most common approach to deriving partial atomic charges is to fit to the electrostatic potential of a molecule as obtained from a QM calculation. However, while the electrostatic potential around an atom can be calculated to high precision, the net charge on an atom is not itself an observable property. As a consequence, charges proposed on the

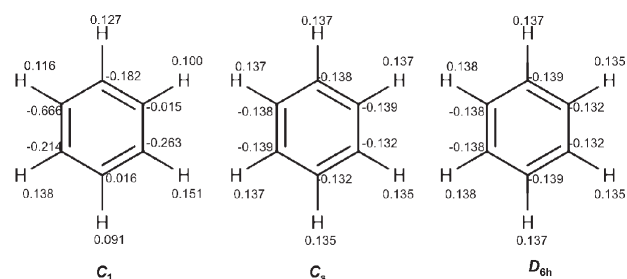


Figure 1. The partial atomic charges derived by fitting to the electrostatic potential at B3LYP/6-31G* level of theory for benzene optimized assuming C_1 , C_s , and D_{6h} symmetry using the method of Kollmann and Singh.³⁹

basis of alternative charge assignment methods can differ significantly.³⁸ In addition, QM-derived charges depend on the level of theory used in the calculations, the precise conformation of the molecule, and the symmetry (point group) used when performing the calculations on the molecule. These differences can be significant even in very simple systems. Figure 1 shows the partial atomic charges obtained by fitting to the electrostatic potential of a molecule of benzene in implicit water calculated using the method of Kollmann–Singh³⁹ starting from the optimized geometry at the B3LYP/6-31G* level of theory as a function of the molecular symmetry (point group). As can be seen, there are dramatic differences in the charges assigned to the atoms using C_1 as opposed to enforcing D_{6h} symmetry. Symmetry must also be considered in the assignment of atom types and bonded parameters if the topologies are to be used in simulations where the molecules can change conformation. Last but not the least, the assignment of bonded parameters also represents a major challenge. Most biomolecular force fields contain bonded parameters for only a narrow range of molecular architectures, and new parameters must be derived for novel cases. However, the actual value of a bond angle in an optimized geometry may differ significantly from the relevant ideal value due to the effect of the local environment captured by other terms in the force field. Likewise, while force constants for specific internal degrees of freedoms can be derived from the Hessian (the derivative of the force with respect to the coordinates) as obtained from QM calculations, these will also contain contributions due to non-bonded interactions within the molecule.

ATB PIPELINE

A flow-chart outlining the steps taken in the generation of a molecular topology using the ATB is outlined in Figure 2. The user is required to provide three pieces of information:

- A set of three-dimensional coordinates of all atoms (including all hydrogen atoms) in PDB format
- A set of atomic connectivity data (PDB CONECT records)
- The overall formal charge on the molecule

The combination of coordinates, connectivity, and overall charge provided by the user uniquely defines the stereochemistry as well as the protonation and tautomeric state of the molecule. In addition, the user is requested to provide a common name or a description of the molecule to assist when searching the database. The user is also asked to provide the IUPAC name of the molecule, any experimental data related to the free energy of hydration of the molecule, and a unique three to four-letter/-digit

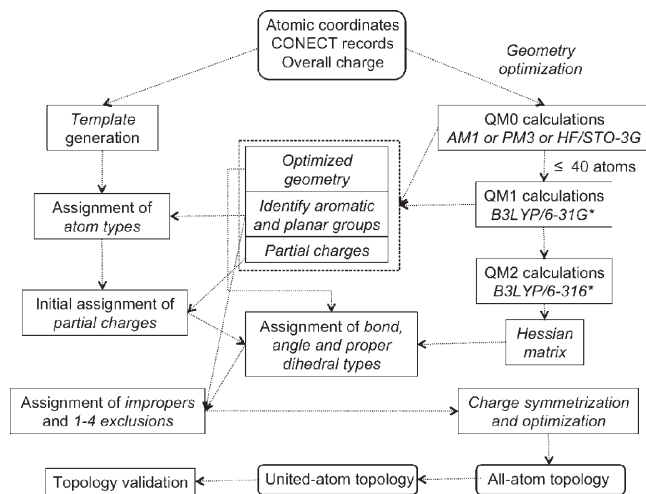


Figure 2. A flow-chart summarizing the primary steps in the generation of a molecular topology using the Automated Topology Builder (ATB).

residue name (RNME). If not provided by the user, a unique residue name will be generated automatically, and the IUPAC name is obtained using the “IUPAC name generator” utility of the program Jchem 5.2.03.1 (2009, ChemAxon, <http://www.chemaxon.com>). The user can also assign the molecule to a particular class: amino acid, nucleic acid, lipid, sugar, solvent, or heteromolecule. The topology builder uses a knowledge-based approach in combination with QM calculations to select parameters consistent with a given version of the GROMOS force field. An important feature of the ATB is that alternate bonded parameters are listed as comments in cases where the assignment is ambiguous. The user can then select the most appropriate parameter from the alternative choices.

The topology is constructed as follows:

1. Generation of the Initial Template. The first stage involves the generation of an initial template for the topology. This initial template contains only information that can be assigned unambiguously on the basis of the sequence of the atoms and the CONECT records. It does not incorporate information based on external rules or information extracted from the QM calculations. The standard GROMOS topology building block file contains various blocks. The TITLE, TOPPHYSCON, LINKEXCLUSIONS, and MTBUILDBLSOLUTE blocks are common to all molecules. The TITLE block can contain arbitrary text. Within the TITLE block, the ATB provides the date and time the file was generated, the IUPAC name of the molecule, a description of the molecule, and the revision date of the ATB program. The ATB also provides additional information as comments. These are placed before the TITLE block in the topology file. This includes general information regarding the ATB, warnings pertaining to any parameters that could not be assigned by the ATB, information on how to choose appropriate type codes when alternate bonded parameters are listed, the level of QM calculations performed, and the method used to calculate the initial partial atomic charges. The TOPPHYSCON block contains a set of standard physical constants. The LINKEXCLUSIONS block contains information on which atoms are to be excluded when linking individual monomeric units when building a polymeric chain in certain versions of the GROMOS force field, for example, when building a protein from a set of amino acid building blocks. Since the ATB generates topologies for complete molecules, the LINKEXCLUSIONS

block is not used. The number of exclusions has been set arbitrarily to 2.

The MTBUILDBLSOLUTE block contains information concerning the molecule itself. The first line contains a unique identifier residue name (RNME). If the user does not supply a unique name, one will be generated by the ATB. This is followed by the number of atoms (NMAT) and the number of preceding exclusions (NLIN). NMAT is derived from the PDB file and is set equal to the number of lines containing atomic coordinates. NLIN is set to zero. NLIN is relevant only when linking more than one monomer in a polymeric chain. Various parameters are then given for each atom in the molecule. These include the atomic serial number (ATOM), the atom name (ANM), the atom type code (IACM), the atomic mass type code (MASS), the partial atomic charge (CGM), the charge group (ICGM), the number of exclusions (MAE), and a list of the excluded atoms (MSAE). ATOM and ANM are derived from columns 7–11 and 13–16 of the PDB coordinate file, respectively. The assignment of IACM depends on the chemical environment and is done at a later stage. The MASS types are assigned on the basis of the elemental symbol in columns 77–78 of the PDB file. The corresponding mass information is taken from the MASSATOMTYPECODE block of the appropriate interaction parameter file (e.g., IFP53A6.dat of the GROMOS 53A6 force field). As no charge has yet been assigned, CGM and ICGM are set to zero. In the GROMOS force field, first and second covalently bound neighbors (1–2 and 1–3 bonded pairs) are normally excluded from nonbonded interactions. In certain cases, such as within aromatic rings, third neighbors (1–4 bonded pairs) are also excluded. These are assigned later. Initially, all possible 1–2 and 1–3 pairs are identified on the basis of the CONECT records of the PDB file and used to assign the exclusions (MAE and MSAE).

The MTBUILDBLSOLUTE block then provides information on the bonds, bond angles, and the proper and improper dihedral angles. The total number of bonds (NB) is given followed by the list of the atom pairs (IB and JB), each with a corresponding bond type (MCB). The number of bonds and the pairs IB and JB are taken directly from the CONECT records. The MCB codes are assigned later. The CONECT records are also used to generate a list of all possible bond angles and proper dihedral angles. From this list, the total number of bond angles (NBA) and the atoms (IB, JB, and KB) forming a given angle are populated. Likewise, the total number of dihedral angles (NDA) and atoms forming the dihedral (IB, JB, KB, and LB) are populated. The corresponding angle and dihedral type codes (MCB) are not assigned at this stage. The assignment of improper dihedrals requires knowledge of whether certain atoms are part of aromatic rings or planar groups and are not assigned at this stage.

2. Quantum Mechanical Calculations. To assist in the assignment of appropriate parameters, a series of QM calculations is performed. All QM calculations are performed using GAMESS-US.⁴⁰ The molecule is initially optimized at the HF/STO-3G, AM1,²¹ or PM3⁴¹ level of theory as selected by the user. If the molecule contains ≤ 40 atoms, it is further optimized at the B3LYP/6-31G* level of theory^{42–44} in an implicit solvent described using the polarizable continuum model (PCM⁴⁵). The Hessian of the system is also calculated.

A range of information is extracted from the output of the QM calculations. This includes the following:

- (i) **Bonded parameters.** The optimized geometry is used to obtain an indication of the bond lengths, bond angles, and

dihedral angles, which are used to help assign appropriate bonds, bond angles, and dihedral type codes (MCB). In addition, the optimized geometry is used to identify any chiral centers within the molecule. The “bond order” is also extracted and used to verify the CONECT records supplied by the user.

(ii) **Identification of aromatic rings and planar groups.**

These are identified from the optimized geometry and assist with the assignment of the atom types and improper dihedral angles along with the relevant 1–4 exclusions. To identify the aromatic rings, all of the rings in a given molecule are first identified on the basis of the loop closure from the connectivity records. Fused rings are identified from the shared connectivity between two or more rings. The rings are initially analyzed separately. Any redundancies are then removed so that the fused system is treated as a single unit. All possible proper dihedral angles that involve bonds that form the ring are identified. If all of these dihedral angles have a value between -5.0° and $+5.0^\circ$, then the ring is classified as an aromatic ring. Next, the atoms (e.g., C), which are connected to three other atoms (e.g., X, O, and N) and are not a part of an aromatic ring, are identified. The improper dihedral between these four atoms as C–X–O–N is determined. If the value is between -10.0° and $+10.0^\circ$, then the group is classified as planar (e.g., an amide group).

(iii) **Initial partial atomic charges.** The initial charges are estimated using either the MOPAC^{46,47} method (for AM1 and PM3) or the method of Mulliken⁴⁸ (for HF/STO-3G) for molecules containing >40 atoms. For the molecules with ≤ 40 atoms, the initial atomic charges are generated by fitting the electrostatic potential (at the B3LYP/6-31G* level of theory) using the Kollmann–Singh³⁹ scheme. These are then further optimized by ATB.

(iv) **The Hessian.** The Hessian is used to estimate harmonic force constants for bond stretching and angle bending degrees of freedom as described by Seminario.⁴⁹

3. Initial Assignment of Atom (IACM) Types. In the GROMOS force field, the IACM type code determines the Lennard-Jones parameters used to model van der Waals interactions. As in most biomolecular force fields, the appropriate IACM depends on the local chemical environment. For example, oxygen, nitrogen, carbon, and hydrogen can be modeled using multiple, different atom types. Thus, while the element type can be assigned on the basis of the elemental symbol in columns 77–78 of the PDB file, the value of IACM must be inferred on the basis of the connectivity, the chemical environment, and in some cases the net charge on the atom. Possible atom type codes are first extracted from the IAC block of the interaction parameter file (e.g., IFP53A6.dat). The GROMOS 53A6 force field contains 53 unique atom types. For oxygen (O), nitrogen (N), carbon (C), and hydrogen (H), the initial assignment is based on the following rules.

For oxygen, there are five different atom types (1–5). Type 1 is assigned in cases where the oxygen is bound to just one C atom (i.e., a carbonyl group; O=C) and where no other rule applies. Currently, type 2 is assigned in all cases where more than one O atom is attached to the same atom but which themselves are not attached to another atom (e.g., an O atom with double bonds in $-\text{CO}_2^-$, $-\text{PO}_4^{-2}$, $-\text{SO}_4^{-2}$, $-\text{SO}_2$, $-\text{NO}_2$, etc. groups). This is appropriate in cases where there is a large partial (negative) charge on the oxygen. In cases where the partial charge is small,

type 1 may be more appropriate. Type 3 is assigned in cases where the oxygen is bound to two atoms, at least one of which is not carbon (e.g., C–O–H, H–O–P, etc.). Type 4 is assigned in cases where the oxygen atom is connected to two carbon atoms [e.g., an O atom with two single bonds in ester or ether; C(=O)–O–C]. Type 5 is specific for the oxygen in a SPC water molecule and is not used for heteromolecules. There are six atom types for nitrogen (6–11). Nonaromatic nitrogen atoms connected to three neighbors are assigned either types 6 or 7. Cases where the nitrogen atom is connected to at most one hydrogen atom and a carbonyl carbon are assigned to type 6 [e.g., a N in $-\text{CONHCH}_3$]. Other nitrogen atoms connected to three neighbors are assigned to type 7 (e.g., a N in $-\text{CONH}_2$, $-\text{NH}_2$, etc.). Nitrogen atoms bound to four other atoms (e.g., a quaternary N atom with +1 charge) are assigned to type 8. When the nitrogen atom forms part of an aromatic ring with either two or three bonded neighbors, it is assigned to type 9. Nitrogen atoms that have a single neighbor (e.g., N in $-\text{C}\#\text{N}$) are also assigned to type 9. The atom types 10 and 11 are specific for the side chain of arginine and are currently not assigned by ATB for heteromolecules. There are eight different atom types for carbon (12–19). Initially, all carbons are assigned to type 12 except if the atom in question is attached to four non-hydrogen atoms. In this case, the atom is assigned to type 13. Types 14–19 correspond to united-atom carbons: CH1, CH2, CH3, CH4, and CH-aromatic. These will be discussed later in relation to the conversion of an all-atom topology into a united-atom topology. Finally, there are two types for hydrogen. Type 20 is used for all hydrogens bound to a carbon, and type 21 is used for all other hydrogen atoms. The other atom type codes either are unique or refer to atoms parametrized for a given solvent. The solvent atom types are not used by ATB, except atom type 42 (sulfur in DMSO solvent in GROMOS), which is used for molecules containing S with more than two connected neighbors.

4. Assignment of Bond and Bond Angle Types (MCB). The bond and bond angles are derived from the BONDTYPECODE and the BONDANGLETYPECOD blocks of the IFP53A6.dat file. The value of the bond, bond angle, and corresponding force constant along with the atom types are matched to the standard GROMOS bond and angle types in the IFP53A6.dat file. The matching threshold is set to ± 0.004 nm for the bond length and $\pm 5.0^\circ$ for the bond angle. For molecules with ≤ 40 atoms, the Hessian is available from the QM calculations, from which the force constants for the bonds and bond angles can be estimated using the method of Seminario.⁴⁹ The matching threshold for the force constant for the bonds is $\pm 1 \times 10^6$ kJ mol⁻¹ nm⁻⁴, while that for bond angles is ± 100.0 kJ mol⁻¹. Note, in the GROMOS 53A6 parameter set, the bond term is quartic in the bond length and the angle term is dependent on the cosine of the bond angle. In cases where a suitable match is not found, a new bond or bond angle type is introduced with the corresponding force constant and value as obtained from the QM calculations. These new parameters are marked as “nonstandard” in the topology file to indicate them appropriately. The angle type 41 in GROMOS is specific for the heme with C#O (Fe–C#O, 180°) and is not used for the heteromolecules. Instead a new (nonstandard) angle type 55 has been introduced with a force constant of 500.0 kJ mol⁻¹ and angle of 180° for molecules with linear groups $-\text{C}\#\text{N}$, $-\text{C}\#\text{C}-$, $-\text{C}=\text{C}=\text{C}-$, $-\text{N}=\text{C}=\text{S}$, etc.

5. Assignment of Proper Dihedral Angles. The proper dihedral angles are derived from the DIHEDRALTYPECODE block of the IFP53A6.dat file. Initially, to build the template

building block, all possible dihedral angles for a given bond are listed. Redundant proper dihedrals are then removed. In most cases in the GROMOS force field, only one set of atoms i, j, k , and l is chosen to define a dihedral angle around the central bond between atoms j and k .

The multiplicity (m_n) of a dihedral angle is determined based on the connectivities of atoms j and k as follows. First, the value of M is determined on the basis of the connectivity information:

$$M = A \times B \quad (1)$$

where $A = (\text{number of atoms connected to } j) - 1$ and $B = (\text{number of atoms connected to } k) - 1$.

The possible values for M in heteromolecules are 1, 2, 3, 4, 6 and 9:

if $M = 1, 2, 3$, or 6, then $m_n = M$

if $M = 4$ or 9, then $m_n = M^{1/2}$, i.e., $m_n = 2$ or 3, respectively.

Thus, possible values of m_n are 1, 2, 3, and 6. The GROMOS force field also contains a dihedral angle with a multiplicity of 4.

This is used specifically for heme groups and is not assigned in the heteromolecules.

Next, the phase shift is determined by evaluating the potential energy of the dihedral angle using eq 2:

$$V^{\text{trig}}(r; s) = V^{\text{trig}}(r, K_{\varphi}, \delta_m) = \sum_{n=1}^{N_{\varphi}} K_{\varphi, n} [1 + \cos(\delta_n) \cos(m_n \varphi_n)] \quad (2)$$

where δ_n is the phase shift, which is restricted to 0 or π (i.e., $\cos \delta_n = \pm 1.0$), m_n is the multiplicity of the torsion dihedral angle, and φ_n is the actual value of the dihedral angle defined by atoms i, j, k , and l . Equation 2 indicates that a phase shift of 180° and a multiplicity of 1 means that the potential reaches the maximum value at 180° (if multiplicity = 3, then at $-60^\circ, 60^\circ$, and 180°). Equation 2 is evaluated for $\cos \delta_n = +1$ or -1 for each dihedral angle in a given molecule with a given multiplicity (m_n) and the dihedral angle value (φ_n). The $\cos \delta_n$ value that gives lower value for the potential energy is taken as the phase shift (δ_n) for that dihedral angle. Possible dihedral type codes for a given dihedral angle are identified by matching the multiplicity and phase shift to types in the IFP53A6.dat file.

The final selection is then based on the combination of atom types. In ambiguous cases, multiple options are presented. In case no suitable match is found, the standard GROMOS dihedral type with the lowest force constant with the corresponding phase shift and multiplicity is chosen.

Dihedral type 16 (force constant of 0.0 kJ mol^{-1}) is not used for heteromolecules. Since there is no dihedral angle type with a phase shift of -1 and multiplicity of 3 in the GROMOS force field, a new (nonstandard) type 42 was introduced with a force constant of 1.0 kJ mol^{-1} . No dihedral is assigned in cases where one of the bond angles involved is 180° , i.e., in molecules containing a linear group such as an alkyne ($-C\#C-$) or azido ($-N\equiv N\equiv N$).

6. Assignment of Improper Dihedral Angles and 1,4 Exclusions in the Case of Aromatic Rings, Planar Groups, and Chiral Centers. Aromatic rings, planar groups, and chiral centers are identified on the basis of the QM optimized geometry as described previously. The improper dihedral angles are assigned from the IMPDIHEDRALTYPE block of the IFP53A6.dat file. To maintain the planarity of an aromatic ring and/or planar groups such as $-\text{NO}_2$ or $-\text{NH}_2$, a type 1 (planar) improper dihedral is assigned. In cases where the chirality of a group cannot be maintained by angle terms alone, such as in the case of a chiral

united-atom $-\text{CH}$ group, a type 2 (tetrahedral) is assigned. Where a proper dihedral angle involving the central two atoms had been assigned previously, such as for atoms involved in an aromatic ring, the proper dihedral is removed and only the equivalent improper dihedral retained. The number of proper dihedrals (NDA) is adjusted accordingly. For those atoms that either form part of an aromatic ring or are attached directly to the ring, all possible 1–4 pairs are determined. These 1–4 pairs are added to the list of exclusions in MAE and the value of MSAE updated.

7. Optimization of the Partial Atomic Charges and Assignment of Charge Groups (CGM and ICGM). The initial partial atomic charges are extracted from the output of the QM calculations, rounded to three decimal places, and put in the column CGM. As discussed previously, the assignment of partial charges to atoms represents a major challenge in force field development. This is because it is not possible to relate the charge on an atom to a physical observable, and while a number of models have been developed that can be used to infer charges based on the electron density as derived from QM calculations, they are based on subjective assumptions and do not yield a unique answer. In addition, the charges derived from QM calculations are often very sensitive to the geometry of the molecule. For example, the charges assigned to equivalent chemical groups in a molecule can differ significantly, making such QM charges inappropriate for use in molecular dynamics simulations. In an attempt to overcome these limitations and obtain charges compatible with the GROMOS force field, a number of charge optimization steps are performed. First, the molecule is analyzed in order to identify any global and/or local symmetry. The charges on atoms in equivalent chemical environments are then averaged appropriately. The effect of this procedure is illustrated in Figure 3, which shows the initial QM derived charges calculated by fitting them to the electrostatic potential using the Kollmann–Singh scheme at the B3LYP/6-31G* level of theory and the final optimized ATB charges for hydroquinone and isopropanol. Hydroquinone contains two sets of equivalent $-\text{O}-\text{H}$ groups and four sets of equivalent $-\text{C}-\text{H}$ groups. These symmetric groups were identified, and the charges on individual atoms were averaged and reassigned (Figure 3a). Isopropanol has both local (within the individual methyl groups) as well as global symmetry (the two equivalent methyl groups). First, the charges on the three equivalent hydrogen atoms of the methyl group are averaged and reassigned. Then, the methyl groups are identified as equivalent and the charge of the corresponding atoms averaged and reassigned (Figure 3b).

Next, charge groups are assigned by grouping small numbers of covalently bound atoms together such that the overall charge in a given charge group is either 0.0 or $+1.0$ or -1.0 . For this, all of the hydrogen atoms attached to a single heavy atom are first merged into a single charge group. Any heavy atoms connected to only one other heavy atom are then identified and merged into the adjoining charge groups. The charge groups containing a residual charge are identified, and the residual charge is transferred to the atom within the charge group with the highest charge such that now the overall charge in the charge group is either 0.0 or $+1.0$ or -1.0 . Next, the total charge of the molecule is calculated and compared to the value supplied by the user. Any residual charge that results from rounding etc. is transferred to the atom with the largest charge. Note, members of a charge group must be numbered sequentially, and thus the atoms in the topology file and the associated coordinate file are reordered if

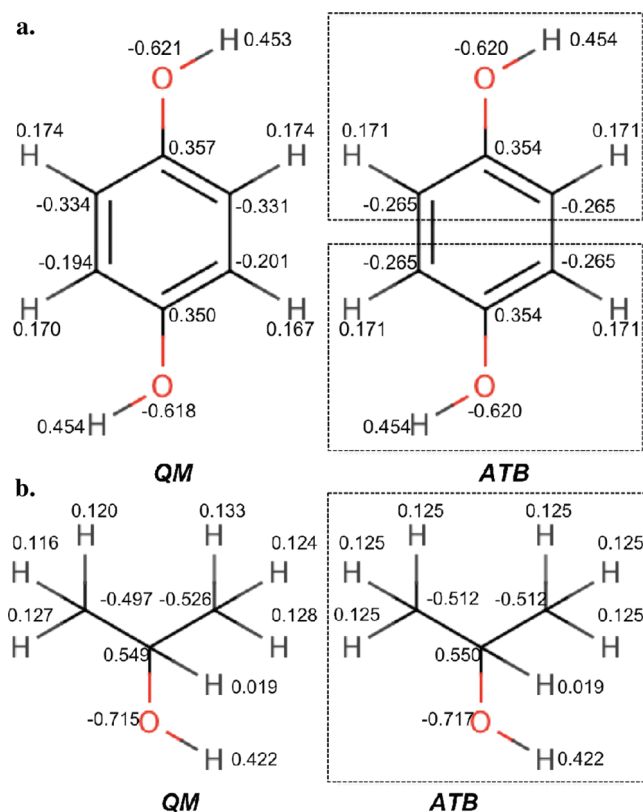


Figure 3. The partial atomic charges derived for molecules (a) hydroquinone and (b) isopropanol from the QM calculations by fitting to the electrostatic potential at the B3LYP/6-31G* level of theory using the method of Kollmann–Singh³⁹ and final ATB charges, respectively. The atoms within the square formed by dotted lines form a single charge group with overall charge of the charge group being either 0.0 (in these cases) or +1.0 or –1.0.

required. This charge optimization process results in atomic charges that are less sensitive to the geometry of the molecule used to calculate the QM charges. The charges are thus more robust and more in line with the philosophy of the GROMOS force field.

8. Conversion from an All-Atom to a United-Atom GROMOS Topology. In the GROMOS force field, aliphatic (alkanes: –CH, –CH₂, –CH₃ and –CH₄) carbon atoms are treated as united atoms. In ATB, only aliphatic carbon atoms are converted to a united-atom model. The alkene (e.g., –CH=CH–), alkyne (e.g. –C#C–), and aromatic –CH groups are retained as all-atom.

A united-atom GROMOS topology building block is generated from the all-atom topology building block in a stepwise manner.

- The C atoms attached to four different atoms and at least one H atom is identified from the CONECT records as either –CH, –CH₂, or –CH₃.
- The partial atomic charges of the H atoms are added to the charge of the corresponding C atom.
- The atom type for each of these C atoms is changed from type 12 to the appropriate united-atom C type, 14, 15, and 16 for –CH, –CH₂, and –CH₃, respectively.
- The H atoms that were attached to these groups are deleted from the topology file and the coordinate file and the remaining atoms renumbered accordingly.
- For a united atom –CH, the improper dihedral angle between –C(X)(Y)Z is calculated. If the improper di-

hedral angle is $35.0^\circ \pm 5.0^\circ$, then an additional improper dihedral angle with type 2 is added to the topology to maintain the chirality of the –CH atom. If the improper dihedral angle is $-35.0^\circ \pm 5.0^\circ$, then one of the pairs from X–Y–Z is swapped (for, e.g., –C(X)(Z)Y) to obtain the value of $35.0^\circ \pm 5.0^\circ$, and the improper dihedral is assigned appropriately. The number of improper dihedrals (NIDA) is updated. This is done for all of the –CH united atoms in the molecule. Note in the case of aliphatic and alicyclic amines, no improper dihedral is assigned, so that pyramidal inversion of the nitrogen atom is possible.

9. Validation of the Force Field Topology. The united-atom GROMOS building block file is validated using the program *check_top* in the GROMOS package.⁵⁰ First, a complete force field topology is constructed from the building block file using the program *make_top*. It is this file that is used in the validation. The program *check_top* performs basic checks on the charge groups, exclusions, bonds, angles, and improper dihedrals. For example, *check_top* will identify charge groups with a residual charge; missing dihedral angles, if the overall charge on the molecule is not a whole number; and possible inconsistencies in the exclusions. If the molecule contains 1–4 exclusions, a warning to check for the presence of aromatic rings in the molecule is issued. Any inconsistencies between the atom type and corresponding charge and bonded types assigned by the ATB and those in the standard GROMOS force field along with possible alternatives are then listed. Finally, the potential energy of each of the bonded terms (bonds, bond angles, proper, and improper dihedrals angles) is then calculated on the basis of the optimized geometry. This allows the user to identify whether a given set of parameters is incompatible with the optimized geometry.

10. Conversion to Other Formats. The final GROMOS force field files can in principle be converted to a wide variety of other formats. Currently, force field files that can be read by the GROMACS simulation package and the CNS structural refinement program are provided. The force field (molecule.itp) files provided for use with GROMACS retain all of the information contained in the original GROMOS file. Information concerning 1–4 pairs for which special van der Waals parameters are assigned as part of the GROMOS force fields are included in the “[pairs]” block in the GROMACS molecule.itp file. In GROMACS, exclusions are generated automatically on the basis of the parameter “*nrexcl*”. By setting *nrexcl* = 3, the 1–2 and 1–3 exclusions are generated automatically. Additional 1–4 exclusions, such as those found in aromatic rings, are included in the “[exclusions]” block. In the molecule.itp file generated by ATB, the parameters for the bonded terms are included explicitly for each interaction to avoid possible incompatibilities when using type codes in GROMACS.

The CNS output contains two files: a topology file and a parameter file. Both are derived from the GROMOS building block file. The topology file contains atom type information. In addition, all bonds, bond angles, and dihedral angles are listed using atom names rather than atom serial numbers. The parameter file contains the interaction parameters. Again, specific values are listed for each of the bonded terms.

REPOSITORY FOR GROMOS FORCE FIELDS AND PRE-EQUILIBRATED SYSTEMS

The ATB also provides a repository for the topologies of molecules that have been individually optimized as part of

the generation of the GROMOS family of force fields, including the GROMOS 43, 45, and 53 parameter sets. The original force field files can be downloaded from “the GROMOS force field” tab on the Web server. A range of equilibrated systems are also provided under the “pre-equilibrated systems” tab. These include a range of common solvents such as methanol, ethanol, and acetone as well as a series of small organic molecules including ethers, esters, alcohols, ketones, and carboxylic acids that were used in the parametrization of the GROMOS 53A6 force field. The pre-equilibrated boxes are available in PDB and GROMOS96 file formats along with the corresponding GROMOS input file, force field topologies in various formats, and details of the simulations (temperature, pressure, step size, time duration for equilibration). Links to relevant literature references are also provided. A range of pre-equilibrated mixed systems such as lipid bilayers is currently being added.

CURRENT LIMITATIONS AND ONGOING DEVELOPMENTS

While the ATB aims to generate topologies for a wide range of heteromolecules, it does have several limitations:

- (i) The GROMOS force field does not contain parameters for all possible atom types. Atoms such as boron and iodine are not included in the standard GROMOS force field. When novel atom types are encountered, the current version of ATB terminates after the generation of the initial template.
- (ii) The GROMOS force field is primarily intended for biomolecular systems. The rules that have been developed to assign specific atom types are limited to those that can be derived from biomolecules present in the current GROMOS force fields.
- (iii) The size of the molecule is limited. Higher-level QM calculations, required to determine the Hessian, are only performed if the molecule contains ≤ 40 atoms.
- (iv) Currently, the Hessian is used only to estimate the force constants for bonds and bond angles. The dihedral angle terms are highly correlated with the nonbonded parameters such as partial charges, Lennard-Jones terms, exclusions, etc. While in principle corrections to the Hessian can be applied, in the current version of the ATB, the force constants for dihedral angles are not estimated on the basis of the Hessian.
- (v) The possibility to scale the partial charges of atoms in specific functional groups in order to provide a better match to the existing GROMOS 53A6 force field has been implemented, but charge scaling is currently not applied. For example, primary amines are known to be problematic in classical force fields, and the magnitude of the optimized charges used in the GROMOS force field are significantly larger than those obtained using QM approaches.⁵¹
- (vi) At present, atom types are assigned on the basis of the local environment as determined by connectivity. The incorporation of atomic charge in the assignment of atom types is under development.
- (vii) Force field topologies are currently only provided in GROMOS, GROMACS, and CNS formats the ATB.

VALIDATION OF THE ATB

To test the validity of the force field descriptions generated by the ATB, the ability to reproduce the thermodynamic properties of a range of molecules has been examined. The GROMOS 53A6 force field has been parametrized to reproduce the density and heats of vaporization of pure organic liquids as well as to reproduce the free energies of solvation of analogs of the side chains of common amino acids in polar and apolar solvents. As a part of the validation, the topologies generated by the ATB have been used to estimate the free energy of hydration of analogs of the side chains of amino acids. In addition, the free energy of hydration of a set of 90 biologically relevant small organic molecules^{52–54} (Table S1, Supporting Information) and 100 chemically diverse drug-like and drug molecules taken from the “Statistical Assessment of the Modeling of Proteins and Ligands” (SAMPL) challenges including the CUP8(SAMPL0),⁵⁵ SAMPL1,⁵⁶ and SAMPL2⁵⁷ data sets has been determined (Table S2, Supporting Information).

As an initial test of the validity of the molecular descriptions generated by the ATB, the all-atom RMSD between the QM optimized structure used in the parametrization and the structure after 1 ns of simulation in SPC water were determined for the 190 test molecules (Tables S1 and S2, Supporting Information). For 156 of these molecules ($\sim 80\%$), the all-atom RMSD between the simulated and QM optimized structure was < 0.1 nm, demonstrating that the QM optimized structure is also a minimum in the force field generated by the ATB. Of the remainder, for 26 molecules, the RMSD was between 0.1 and 0.2 nm, and for eight molecules, the RMSD was above 0.2 nm. All of these molecules, however, contained either an alkane chain or a similar flexible group. Ketoprofen (Table S2), which contains a slightly flexible and large benzophenone ring with a branched chain, exhibited the highest RMSD of 0.27 nm.

The free energies of hydration were calculated using the GROMOS96⁵⁸ simulation package. A given solute molecule was placed in a cubic periodic box filled with SPC⁵⁹ water molecules. The size of the box was chosen such that no solvent molecule interacted with more than one periodic image of the solute. After energy minimization, the initial velocities were assigned from a Maxwell–Boltzmann distribution corresponding to a temperature of 298.15 K. All bond lengths were constrained using the SHAKE⁶⁰ algorithm with a relative geometry accuracy of 10^{-4} . The equations of motion were integrated using the leapfrog algorithm and a time step of 2 fs. All simulations were performed at a constant temperature (298.15 K) and pressure (1 atm) using a Berendsen thermostat (coupling time of 0.1 ps) and barostat (coupling time of 0.5 ps and isothermal compressibility of 4.575×10^{-4} (kJ mol⁻¹ nm⁻³)⁻¹).⁶¹ Nonbonded interactions were calculated using a triple-range scheme. Interactions within a shorter-range cutoff of 0.8 nm were calculated every time step from a pair list that was generated every five steps. At these time points, interactions between 0.8 and 1.4 nm were calculated as well and kept constant between updates. A reaction field contribution was added to the electrostatic interactions and forces, to account for a homogeneous medium outside the long-range cutoff. The relative permittivity for the reaction field was set to a value of 61 for SPC water.

The free energy of solvation was calculated using the thermodynamic integration (TI)⁶² approach. To determine the free energy of hydration, all nonbonded interactions involving solute atoms were scaled to zero in a stepwise manner as a function of a coupling parameter λ . The change in free energy corresponding

Table 1. Comparison between the Experimental (exptl) and Calculated (calcd) Free Energies of Hydration (ΔG_{hyd}) for Analogs of the Side Chains of Common Amino Acids Calculated Using the GROMOS96 Force Field and Parameters Assigned by the Automated Topology Builder (ATB)^a

amino acid	side chain analog	$\Delta G_{\text{hyd;exptl}}^1$	$\Delta G_{\text{hyd;calcd}}$		$ \Delta G_{\text{hyd;calcd}} - \Delta G_{\text{hyd;exptl}} $	
			GROMOS 53A6	GROMOS 53A6	$\Delta G_{\text{hyd;calcd}}$ ATB	$ \Delta G_{\text{hyd;calcd}} - \Delta G_{\text{hyd;exptl}} $ ATB
Ala	methane	8.1; 8.4	6.8	1.3; 1.5	6.8	1.3; 1.5
Arg	<i>n</i> -propyl-guanidine	-45.7	-48.6	2.9	-43.0	2.7
Asn	acetamide	-40.6	-40.9	0.3	-37.9	2.7
Asp	acetic acid	-28.0	-30.5	2.5	-29.8	1.8
Cys	methane thiol	-5.2	-6.3	1.1	-8.4	3.2
Gln	propanamide	-39.4	-40.4	1.0	-34.4	5.0
Glu	propionic acid	-27.0	-30.0	3.0	-31.6	4.6
His	methyl imidazole	-42.9	-44.5	1.6	-42.9	0.0
Ile	<i>n</i> -butane	8.7; 8.8	8.8	0.1; 0.0	7.4	1.3; 1.4
Leu	isobutane	9.4; 9.7	10.0	0.6; 0.3	8.2	1.2; 1.5
Lys	<i>n</i> -butyl-amine	-18.3	-20.0	1.7	-9.0	9.3
Met	ethyl methyl sulfide	-6.2	-7.6	1.4	1.8	8.0
Phe	toluene	-3.1	-1.0	2.1	1.9	5.0
Ser	methanol	-21.2	-23.1	0.9	-21.9	0.7
Thr	ethanol	-20.5	-19.9	0.6	-21.8	1.3
Trp	3-methyl-indole	-24.7	-24.5	0.2	-16.2	8.5
Tyr	<i>p</i> -cresol	-26.6	-25.1	1.5	-25.8	0.8
Val	propane	8.2	7.8	0.4	7.2	1.0
average					1.2	3.3

^a The absolute value of the difference between the calculated and experimental values is also given. Values are in kJ mol^{-1} .

to the removal of all solute nonbonded interactions was then calculated by integrating the average value of the derivative of the Hamiltonian H of the system with respect to λ :

$$\Delta G = \int_0^1 \left\langle \frac{\partial H}{\partial \lambda} \right\rangle_{\lambda} d\lambda \quad (3)$$

The integral above was evaluated using 21 evenly spaced λ points with 1 ns of data collection at each λ point. Standard errors in $\langle \partial H / \partial \lambda \rangle_{\lambda}$ were estimated at every λ point using a block-averaging procedure.⁶³ The individual errors were integrated from $\lambda = 0$ to $\lambda = 1$ to yield the estimate of the total error in ΔG . A soft-core^{64,65} interaction was used to avoid singularities in the nonbonded interaction function at the end state ($\lambda = 1$). The free energy of solvation was calculated as the difference between the free energy change calculated from a vacuum simulation of the solute and the free energy change when the solute is in solution. In the vacuum simulation, a stochastic bath was applied at a reference temperature of 298.15 K using an atomic friction coefficient of 91 ps^{-1} .

The primary target function in the parametrization of the GROMOS 53A6 force field was the ability to reproduce the free energy of hydration of biomolecular systems. Table 1 lists the free energy of hydration for analogs of the side chains of common amino acids calculated using parameters taken from the GROMOS 53A6 force field, calculated using united atom parameters generated using the ATB and measured experimentally. Note, the ATB uses a combination of structure calculations and a rule-based approach to obtain parameters. It aims to be compatible with the GROMOS force field but does not attempt to directly match specific functional groups to those in the GROMOS 53A6 force field. Thus, while atom types and bonded parameters will match exactly, there will (potentially) be differences in the partial charges and charge groups, as in the GROMOS force field these

have been manually optimized individually for each group. Nevertheless, there is a very close correspondence between the values calculated using the GROMOS 53A6 force field and those values calculated using the force field generated using the ATB. This demonstrates that the parameters generated by the ATB are compatible with those of the GROMOS 53A6 force field. Most importantly, the ability of the two sets of parameters to reproduce the experimental free energies of hydration is similar in most cases. The average deviation between the values calculated using the GROMOS 53A6 force field and the experimental values is 1.2 kJ mol^{-1} . The average deviation between the values calculated using the ATB and the experimental values is marginally larger at 3.3 kJ mol^{-1} . In several cases, the difference between the calculated and experimental free energies of hydration is greater for the GROMOS 53A6 force field than for parameters derived from the ATB. As illustrated graphically in Figure 4, which shows the calculated hydration free energies plotted against the values obtained experimentally, the increased average deviation is primarily due to just three compounds *n*-butyl-amine (Lys; dev = 9.3 kJ mol^{-1}), ethyl methyl sulfide (Met; dev = 8.0 kJ mol^{-1}), and 3-methyl-indole (Trp; dev = 8.5 kJ mol^{-1}). Primary amines, aromatic nitrogens, and sulfur-containing groups are known to be problematic within the GROMOS force field, suggesting either that specific scaling factors for the charges in these cases may need to be introduced or that alternative van der Waals parameters need to be developed for these atoms types.

To further validate the parameters generated by the ATB, the free energies of hydration of ~ 90 biologically relevant small organic molecules were calculated. The molecules encompass chemical classes such as alkanes, cycloalkanes, alkenes, alkynes, alkyl benzenes, amines, amides, aldehydes, carboxylic acids, esters, ketones, thiols, and sulfides. The name, access code, and chemical structure of the compounds in the ATB together with the experimental

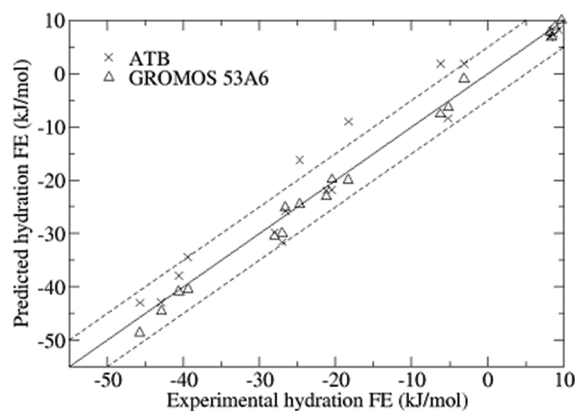


Figure 4. Free energy of hydration of amino acid side chain analogs. Comparison of experimental free energy of hydration to calculated values obtained using the GROMOS 53A6 force field (triangles) and ATB (crosses) for 18 compounds listed in Table 1. The diagonal line corresponds to perfect agreement with experimental results. Dotted lines indicate $\pm 5.0 \text{ kJ mol}^{-1}$ deviations from the diagonal line.

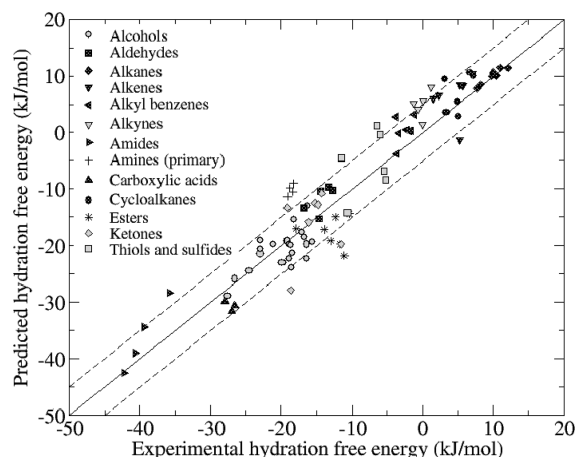


Figure 5. Free energy of hydration of ~ 90 biologically relevant small organic molecules. The diagonal line corresponds to perfect agreement with experimental results. Dotted lines indicate $\pm 5.0 \text{ kJ mol}^{-1}$ deviations from the diagonal line.

and calculated hydration free energies as well as the difference between these values are given in Table S1 as Supporting Information. The average deviation between the values calculated using the force field obtained from the ATB and the experimental values is 3.3 kJ mol^{-1} specifically, alcohols, 2.4 kJ mol^{-1} ; aldehydes, 2.0 kJ mol^{-1} ; alkanes, 0.4 kJ mol^{-1} ; alkenes, 3.9 kJ mol^{-1} ; alkyl benzenes, 3.0 kJ mol^{-1} ; alkynes, 5.0 kJ mol^{-1} ; amides, 3.7 kJ mol^{-1} ; amines (primary), 8.5 kJ mol^{-1} ; carboxylic acids, 3.5 kJ mol^{-1} ; cycloalkanes, 2.5 kJ mol^{-1} ; esters, 5.1 kJ mol^{-1} ; ketones, 2.1 kJ mol^{-1} ; and thiols and sulfides, 5.0 kJ mol^{-1} . In Figure 5, the calculated free energies of hydration are plotted versus the experimental values. While for cases such as simple alkanes the calculated hydration free energies were distributed around the experimental values, the hydration free energies of, in particular, amines (primary), alkynes, and sulfides were systematically overestimated, while esters were systematically underestimated. This suggests that the application of scaling factors on the charges and/or the reparameterization of atoms involved in these specific groups would lead to systematic

improvements in the predictive power of the force field. The magnitudes of the partial atomic charges derived from the QM calculations for sulfides, for example, are significantly smaller than those used for the side chain of methionine in the GROMOS force field. Primary amines and esters are known to be problematic in the GROMOS force field, and work to improve these groups is currently underway.

As a further validation of the ATB, the hydration free energies of ~ 100 drug-like molecules have also been calculated (Supporting Information Table S2). The test molecules were taken from the CUP8(SAMPL0), SAMPL1 and SAMPL2 data sets and represent a very diverse range of functional groups, including groups that do not currently form part of the GROMOS force field and groups that are known to be nonoptimal in the GROMOS force field. Of the 100 molecules, 40 contained at least one halogen ($-\text{F}$, $-\text{Cl}$, $-\text{Br}$). A small number of halogen-containing compounds, notably trifluoroethanol and chloroform, have been specifically parametrized as part of the GROMOS force field. However, the transferability of these parameters has not been investigated systematically. It should also be noted that the uncertainty in the experimental hydration free energies was up to 8.1 kJ mol^{-1} in many cases, much larger than the statistical uncertainty in the calculated values.

Of the 100 molecules, the calculated hydration free energy was within the combined statistical uncertainty of the calculation and experiment in 34 cases. In an additional 25 cases, the calculated value lay within 5 kJ mol^{-1} of the experimental value, again allowing for the combined statistical uncertainty. Overall, the average deviation between the values calculated using the force field obtained from the ATB and the experimental values was 9.1 kJ mol^{-1} . This was primarily due to molecules containing halogens. Of 40 molecules containing one or more halogens, the free energy of hydration was overpredicted by more than 5 kJ mol^{-1} in 28 cases. Furthermore, the extent to which the free energy of hydration was overestimated was strongly correlated with the number of halogens in the molecule unless the molecule contained additional compensating functional groups. The average deviation for the 60 molecules remaining was 7.1 kJ mol^{-1} . In general, the free energies of hydration for molecules containing $-\text{NO}_2$, ether, and/or N-alkyl groups were overestimated, while molecules containing an ester or P or S were underestimated. Again, the systematic nature of these deviations suggests that there is scope for the optimization of these groups, and work to this end is underway.

CONCLUSIONS

The automatic generation of molecular force fields for novel molecules compatible with biomolecular force fields such as GROMOS, AMBER, CHARMM, and OPLS is an ongoing challenge. The Automated force field Topology Builder (ATB, <http://compbio.biosci.uq.edu.au/atb>) and repository described here is a unique Web server that can provide topologies and parameters for a wide range of molecules appropriate for use in molecular simulations, computational drug design, and X-ray refinement. It has the advantage over other comparable servers in that the user is required to supply sufficient information (the coordinates and connectivity of all atoms in the molecule along with the formal charge) in order to define uniquely the geometry and stereochemistry as well as the protonation and tautomeric states of the molecule. The ATB combines information from QM calculations with a knowledge-based approach to derive both

all-atom and united-atom force field descriptions of novel molecules compatible with the GROMOS force field in a variety of formats, including the one for the X-ray refinement program CNS. The symmetry of the molecule is analyzed to ensure that equivalent atoms are assigned identical parameters. Charge groups are assigned automatically. An important feature of the ATB is that it is recognized that it is not possible to unambiguously assign parameters in many cases, and a range of possible alternatives is provided where appropriate. The ATB also acts as a repository for the GROMOS family of force fields and a range of pre-equilibrated systems. At the time of submission, the repository contained an excess of 2100 molecules, including over 100 sugars and 60 lipids.

■ ASSOCIATED CONTENT

S **Supporting Information.** Tables containing the hydration free energy and RMSD data for 190 molecules. This information is available free of charge via the Internet at <http://pubs.acs.org>.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: a.e.mark@uq.edu.au, a.malde@uq.edu.au.

Author Contributions

Scientific aspects and research design: A.K.M. and A.E.M. Code development: A.K.M., L.Z., M.B., M.S., and A.E.M. Manuscript writing: A.K.M. and A.E.M. Validation of the ATB: A.K.M. and A.E.M. Contribution to ATB repository: A.K.M., M.S., D.P., P.C.N., C.O., and A.E.M.

■ ACKNOWLEDGMENT

The authors would like to thank Wilfred van Gunsteren, Philippe Hünenberger, Volker Knecht, Alexandre Bonvin, and Bruno Horta for valuable discussions and Nikhil Biyani for his work on the Hessian. We would also like to thank the many users who have helped identify and correct errors during the development of the ATB. A.K.M. acknowledges the Australian Research Council (ARC; grant no. DP0987043) for the award of an Australian Post-Doctoral (APD) fellowship. P.C.N. acknowledges the Australian government and the University of Queensland for the award of an Endeavour International Postgraduate Research Scholarship (IPRS). Computational resources were provided by the National Computational Infrastructure (NCI, Australia) National Facility through projects m72 and n63.

■ REFERENCES

- (1) Oostenbrink, C.; Villa, A.; Mark, A. E.; van Gunsteren, W. F. A biomolecular force field based on the free enthalpy of hydration and solvation: The GROMOS force-field parameter sets 53A5 and 53A6. *J. Comput. Chem.* **2004**, *25*, 1656.
- (2) Soares, T. A.; Daura, X.; Oostenbrink, C.; Smith, L. J.; van Gunsteren, W. F. Validation of the GROMOS force-field parameter set 45A3 against nuclear magnetic resonance data of hen egg lysozyme. *J. Biomol. NMR* **2004**, *30*, 407.
- (3) Lins, R. D.; Hünenberger, P. H. A new GROMOS force field for hexopyranose-based carbohydrates. *J. Comput. Chem.* **2005**, *26*, 1400.
- (4) Oostenbrink, C.; Soares, T. A.; van der Vegt, N. F. A.; van Gunsteren, W. F. Validation of the 53A6 GROMOS force field. *Eur. Biophys. J. Biophys. Lett.* **2005**, *34*, 273.

- (5) Soares, T. A.; Hünenberger, P. H.; Kastenholz, M. A.; Krautler, V.; Lenz, T.; Lins, R. D.; Oostenbrink, C.; van Gunsteren, W. F. An improved nucleic acid parameter set for the GROMOS force field. *J. Comput. Chem.* **2005**, *26*, 725.
- (6) Jayaram, B.; Sprous, D.; Beveridge, D. L. Solvation free energy of biomacromolecules: Parameters for a modified generalized born model consistent with the AMBER force field. *J. Phys. Chem. B* **1998**, *102*, 9571.
- (7) Wang, J. M.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and testing of a general amber force field. *J. Comput. Chem.* **2004**, *25*, 1157.
- (8) Patel, S.; Brooks, C. L., III. CHARMM fluctuating charge force field for proteins: I parameterization and application to bulk organic liquid simulations. *J. Comput. Chem.* **2004**, *25*, 1.
- (9) Patel, S.; Mackerell, A. D.; Brooks, C. L., III. CHARMM fluctuating charge force field for proteins: II - Protein/solvent properties from molecular dynamics simulations using a nonadditive electrostatic model. *J. Comput. Chem.* **2004**, *25*, 1504.
- (10) Jorgensen, W. L.; Tirado-Rives, J. The OPLS potential functions for proteins - Energy minimizations for crystals of cyclic-peptides and crambin. *J. Am. Chem. Soc.* **1988**, *110*, 1657.
- (11) 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.
- (12) Allinger, N. L. Conformational-analysis 0.130. MM2 - Hydrocarbon force-field utilizing V1 and V2 torsional terms. *J. Am. Chem. Soc.* **1977**, *99*, 8127.
- (13) Allinger, N. L.; Yuh, Y. H.; Lii, J. H. Molecular mechanics - The MM3 force-field for hydrocarbons. 1. *J. Am. Chem. Soc.* **1989**, *111*, 8551.
- (14) Allinger, N. L.; Chen, K. S.; Lii, J. H. An improved force field (MM4) for saturated hydrocarbons. *J. Comput. Chem.* **1996**, *17*, 642.
- (15) Allinger, N. L.; Chen, K. H.; Lii, J. H.; Durkin, K. A. Alcohols, ethers, carbohydrates, and related compounds. I. The MM4 force field for simple compounds. *J. Comput. Chem.* **2003**, *24*, 1447.
- (16) Halgren, T. A. Merck molecular force field. 1. Basis, form, scope, parameterization, and performance of MMFF94. *J. Comput. Chem.* **1996**, *17*, 490.
- (17) Halgren, T. A. Merck molecular force field. 2. MMFF94 van der Waals and electrostatic parameters for intermolecular interactions. *J. Comput. Chem.* **1996**, *17*, 520.
- (18) Halgren, T. A. MMFF VII. Characterization of MMFF94, MMFF94s, and other widely available force fields for conformational energies and for intermolecular-interaction energies and geometries. *J. Comput. Chem.* **1999**, *20*, 730.
- (19) Moriarty, N. W.; Grosse-Kunstleve, R. W.; Adams, P. D. electronic Ligand Builder and Optimization Workbench (eLBOW): a tool for ligand coordinate and restraint generation. *Acta Crystallogr., Sect. D* **2009**, *65*, 1074.
- (20) Adams, P. D.; Afonine, P. V.; Bunkoczi, G.; Chen, V. B.; Davis, I. W.; Echols, N.; Headd, J. J.; Hung, L. W.; Kapral, G. J.; Grosse-Kunstleve, R. W.; McCoy, A. J.; Moriarty, N. W.; Oeffner, R.; Read, R. J.; Richardson, D. C.; Richardson, J. S.; Terwilliger, T. C.; Zwart, P. H. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr., Sect. D* **2010**, *66*, 213.
- (21) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. The development and use of quantum-mechanical molecular-models 0.76. AM1 - A new general-purpose quantum-mechanical molecular-model. *J. Am. Chem. Soc.* **1985**, *107*, 3902.
- (22) Bailey, S. The CCP4 Suite - Programs for protein crystallography. *Acta Crystallogr., Sect. D* **1994**, *50*, 760.
- (23) Vagin, A. A.; Murshudov, G. N.; Strokopytov, B. V. BLANC: the program suite for protein crystallography. *J. Appl. Crystallogr.* **1998**, *31*, 98.
- (24) Vagin, A. A.; Steiner, R. A.; Lebedev, A. A.; Potterton, L.; McNicholas, S.; Long, F.; Murshudov, G. N. REFMAC5 dictionary: organization of prior chemical knowledge and guidelines for its use. *Acta Crystallogr., Sect. D* **2004**, *60*, 2184.
- (25) Emsley, P.; Lohkamp, B.; Scott, W. G.; Cowtan, K. Features and development of Coot. *Acta Crystallogr., Sect. D* **2010**, *66*, 486.

- (26) Brunger, A. T.; Adams, P. D.; Clore, G. M.; DeLano, W. L.; Gros, P.; Grosse-Kunstleve, R. W.; Jiang, J. S.; Kuszewski, J.; Nilges, M.; Pannu, N. S.; Read, R. J.; Rice, L. M.; Simonson, T.; Warren, G. L. Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr., Sect. D* **1998**, *54*, 905.
- (27) Brunger, A. T. Version 1.2 of the Crystallography and NMR system. *Nat. Protoc.* **2007**, *2*, 2728.
- (28) Kleywegt, G. J. Crystallographic refinement of ligand complexes. *Acta Crystallogr., Sect. D* **2007**, *63*, 94.
- (29) Nilsson, K.; Lecerof, D.; Sigfridsson, E.; Ryde, U. An automatic method to generate force-field parameters for hetero-compounds. *Acta Crystallogr., Sect. D* **2003**, *59*, 274.
- (30) Malde, A. K.; Mark, A. E. Challenges in the determination of the binding modes of non-standard ligands in X-ray crystal complexes. *J. Comput.-Aided Mol. Des.* **2011**, *25*, 1.
- (31) Ribeiro, A.; Horta, B. A. C.; de Alencastro, R. B. MKTOP: a program for automatic construction of molecular topologies. *J. Braz. Chem. Soc.* **2008**, *19*, 1433.
- (32) Wang, J. M.; Wang, W.; Kollman, P. A.; Case, D. A. Automatic atom type and bond type perception in molecular mechanical calculations. *J. Mol. Graphics* **2006**, *25*, 247.
- (33) 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.
- (34) Jakalian, A.; Jack, D. B.; Bayly, C. I. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation. *J. Comput. Chem.* **2002**, *23*, 1623.
- (35) Miller, B. T.; Singh, R. P.; Klauda, J. B.; Hodoscek, M.; Brooks, B. R.; Woodcock, H. L. CHARMMing: A new, flexible web portal for CHARMM. *J. Chem Inf. Model.* **2008**, *48*, 1920.
- (36) Schuttelkopf, A. W.; van Aalten, D. M. F. PRODRG: a tool for high-throughput crystallography of protein-ligand complexes. *Acta Crystallogr., Sect. D* **2004**, *60*, 1355.
- (37) Lemkul, J. A.; Allen, W. J.; Bevan, D. R. Practical Considerations for Building GROMOS-Compatible Small-Molecule Topologies. *J. Chem Inf. Model.* **2010**, *50*, 2221.
- (38) Sigfridsson, E.; Ryde, U. Comparison of methods for deriving atomic charges from the electrostatic potential and moments. *J. Comput. Chem.* **1998**, *19*, 377.
- (39) Singh, U. C.; Kollman, P. A. An approach to computing electrostatic charges for molecules. *J. Comput. Chem.* **1984**, *5*, 129.
- (40) Schmidt, M. W.; Baldrige, K. K.; Boatz, J. A.; Elbert, S. T.; Gordon, M. S.; Jensen, J. H.; Koseki, S.; Matsunaga, N.; Nguyen, K. A.; Su, S. J.; Windus, T. L.; Dupuis, M.; Montgomery, J. A. General atomic and molecular electronic-structure system. *J. Comput. Chem.* **1993**, *14*, 1347.
- (41) Carnall, W. T.; Fields, P. R.; Rajnak, K. Electronic energy levels in trivalent lanthanide aquo ions. I. Pr³⁺ Nd³⁺ Pm³⁺ Sm³⁺ Dy³⁺ Ho³⁺ + Er³⁺ and Tm³⁺. *J. Chem. Phys.* **1968**, *49*, 4424.
- (42) Becke, A. D. Density-functional thermochemistry. 3. The role of exact exchange. *J. Chem. Phys.* **1993**, *98*, 5648.
- (43) Lee, C. T.; Yang, W. T.; Parr, R. G. Development of the correlation-energy formula into a functional of the electron-density. *Phys. Rev. B* **1988**, *37*, 785.
- (44) Perdew, J. P.; Wang, Y. Accurate and simple analytic representation of the electron-gas correlation-energy. *Phys. Rev. B* **1992**, *45*, 13244.
- (45) Miertus, S.; Scrocco, E.; Tomasi, J. Electrostatic interaction of a solute with a continuum. A direct utilization of ab initio molecular potentials for the prevision of solvent effects. *Chem. Phys.* **1981**, *55*, 117.
- (46) Chirgwin, B. H.; Coulson, C. A. The electronic structure of conjugates systems. 6. *Proc. R. Soc. London, Ser. A* **1950**, *201*, 196.
- (47) Stewart, J. J. P. Special issue - MOPAC - a semiempirical molecular-orbital program. *J. Comput.-Aided Mol. Des.* **1990**, *4*, 1.
- (48) Mulliken, R. S. Electronic population analysis on LCAO-MO molecular wave functions. 1. *J. Chem. Phys.* **1955**, *23*, 1833.
- (49) Seminario, J. M. Calculation of intramolecular force fields from second-derivative tensors. *Int. J. Quantum Chem.* **1996**, *60*, 1271.
- (50) Christen, M.; Hunenberger, P. H.; Bakowies, D.; Baron, R.; Burgi, R.; Geerke, D. P.; Heinz, T. N.; Kastenholz, M. A.; Krautler, V.; Oostenbrink, C.; Peter, C.; Trzesniak, D.; van Gunsteren, W. F. The GROMOS software for biomolecular simulation: GROMOS05. *J. Comput. Chem.* **2005**, *26*, 1719.
- (51) Oostenbrink, C.; Juchli, D.; van Gunsteren, W. F. Amine hydration: A united-atom force-field solution. *ChemPhysChem* **2005**, *6*, 1800.
- (52) Gerber, P. R. Charge distribution from a simple molecular orbital type calculation and non-bonding interaction terms in the force field MAB. *J. Comput.-Aided Mol. Des.* **1998**, *12*, 37.
- (53) 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.
- (54) Marenich, A. V.; Kelly, C. P.; Thompson, J. D.; Hawkins, G. D.; Chambers, C. C.; Giesen, D. J.; Winget, P.; Cramer, C. J.; Truhlar, D. G. *Minnesota Solvation Database*, version 2009; University of Minnesota: Minneapolis, MN, 2009.
- (55) Nicholls, A.; Mobley, D. L.; Guthrie, J. P.; Chodera, J. D.; Bayly, C. I.; Cooper, M. D.; Pande, V. S. Predicting small-molecule solvation free energies: An informal blind test for computational chemistry. *J. Med. Chem.* **2008**, *51*, 769.
- (56) Guthrie, J. P. A Blind Challenge for Computational Solvation Free Energies: Introduction and Overview. *J. Phys. Chem. B* **2009**, *113*, 4501.
- (57) Geballe, M. T.; Skillman, A. G.; Nicholls, A.; Guthrie, J. P.; Taylor, P. J. The SAMPL2 blind prediction challenge: introduction and overview. *J. Comput.-Aided Mol. Des.* **2010**, *24*, 259.
- (58) van Gunsteren, W. F.; Billeter, S. R.; Eising, A. A.; Hunenberger, P. H.; Kruger, P.; Mark, A. E.; Scott, W. R. P.; Tironi, I. G. *Biomolecular Simulations: The GROMOS96 Manual and User Guide*; Biomos: Zurich, Switzerland, 1996.
- (59) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Hermans, J. Interaction models for water in relation to protein hydration. In *Intermolecular Forces*; Reidel: Dordrecht, The Netherlands, 1981; p 331.
- (60) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. Numerical-integration of cartesian equations of motion of a system with constraints - Molecular-dynamics of N-alkanes. *J. Comput. Phys.* **1977**, *23*, 327.
- (61) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Dinola, A.; Haak, J. R. Molecular-dynamics with coupling to an external bath. *J. Chem. Phys.* **1984**, *81*, 3684.
- (62) van Gunsteren, W. F.; Beutler, T. C.; Fraternali, F.; King, P. M.; Mark, A. E.; Smith, P. E. *Computer Simulation of Biomolecular Systems, Theoretical and Experimental Applications*; ESCOM Science Publishers: Leiden, The Netherlands, 1989; p 315.
- (63) Allen, M. P.; Tildesley, D. J. *Computer Simulations of Liquids*; Oxford University Press Inc.: New York, 1989.
- (64) Beutler, T. C.; Mark, A. E.; van Schaik, R. C.; Gerber, P. R.; van Gunsteren, W. F. Avoiding Singularities and numerical instabilities in free-energy calculations based on molecular simulations. *Chem. Phys. Lett.* **1994**, *222*, 529.
- (65) Zacharias, M.; Straatsma, T. P.; McCammon, J. A. Separation-shifted scaling, a new scaling method for Lennard-Jones interactions in thermodynamic integration. *J. Chem. Phys.* **1994**, *100*, 9025.

Directional Dependence of Hydrogen Bonds: A Density-Based Energy Decomposition Analysis and Its Implications on Force Field Development

Zhenyu Lu,[†] Nengjie Zhou,[†] Qin Wu,[‡] and Yingkai Zhang^{*,†}

[†]Department of Chemistry, New York University, New York, New York 10003, United States

[‡]Center for Functional Nanomaterials, Brookhaven National Laboratory, Upton, New York 11973, United States

ABSTRACT: One well-known shortcoming of widely used biomolecular force fields is the description of the directional dependence of hydrogen bonding (HB). Here we aim to better understand the origin of this difficulty and thus provide some guidance for further force field development. Our theoretical approaches center on a novel density-based energy decomposition analysis (DEDA) method (*J. Chem. Phys.* **2009**, *131*, 164112), in which the frozen density energy is variationally determined through constrained search. This unique and most significant feature of DEDA enables us to find that the frozen density interaction term is the key factor in determining the HB orientation, while the sum of polarization and charge-transfer components shows very little HB directional dependence. This new insight suggests that the difficulty for current nonpolarizable force fields to describe the HB directional dependence is not due to the lack of explicit polarization or charge-transfer terms. Using the DEDA results as reference, we further demonstrate that the main failure coming from the atomic point charge model can be overcome largely by introducing extra charge sites or higher order multipole moments. Among all the electrostatic models explored, the smeared charge distributed multipole model (up to quadrupole), which also takes account of charge penetration effects, gives the best agreement with the corresponding DEDA results. Meanwhile, our results indicate that the van der Waals interaction term needs to be further improved to better model directional HB.

1. INTRODUCTION

While molecular modeling based on molecular mechanical force fields is becoming an indispensable tool in studying structural and dynamical properties of biomolecular systems, it has long been recognized that its applicability and reliability are critically dependent on the accuracy of the employed force field.^{1–5} One main concern for currently widely employed biomolecular force fields, such as CHARMM,⁶ OPLS-AA,⁷ and AMBER,^{8,9} is the description of the directional dependence of hydrogen bonding at the receptor atom.^{1,10–17} For a hydrogen bond D–H···A, this refers to the approaching direction of the hydrogen atom to the acceptor atom A in relation to the bond(s) that A has,^{10,18} as illustrated in Figure 1. Classical force fields employing atomic point charge models and Lenard-Jones (LJ) potentials would lead to very different directional preferences in comparison with results from both high-level quantum chemical calculations of model complexes and detailed analyses of crystal structures.¹⁰ This inadequacy in describing hydrogen-bond (HB) directional dependence has been attributed as one main factor that limits the accuracy and predictive power of force fields in modeling hydrogen-bonding systems,^{1,10–17} including peptide conformation preference,^{11,15,19} protein folding,^{1,14,17} protein–protein interactions,^{12,13} and ligand binding specificity.²⁰

Regarding the origin of hydrogen-bonding directionality, it is widely believed that in addition to the electrostatics, the HB directionality also critically depends on the polarization interactions and the charge transfer from the lone pair (n) of the HB acceptor to the antibonding orbital (σ^*) of the HB donor.^{1,10–17} This opinion has been well supported by molecular orbital-based

energy decomposition analysis (EDA) of intermolecular interactions.^{18,21,22} Currently, there are in general three strategies to tackle this HB directionality problem: (1) The addition of an explicit angle-dependent hydrogen-bonding term to take account of the charge-transfer effect, in which new parameters can be derived from the analysis of protein structure database^{12,19} or fitted to ab initio quantum mechanics (QM) calculations;^{16,23,24} (2) going beyond the atomic point charge model by introducing off-center charges to mimic lone pair (LP) electrons^{25–30} or employing high-order distributed multipoles to better describe electrostatics;^{31–33} and (3) introduction of explicit polarization with induced point dipole,^{25,34–37} drude oscillator,^{38–43} or fluctuating charge models.^{44–47} In spite of substantial efforts, it remains a significant challenge to fundamentally understand this HB directionality problem.

Energy decomposition analysis (EDA)^{48–55} based on high-level quantum mechanical calculations is a powerful tool to study and analyze intermolecular interactions and has been increasingly playing an important role in force field development.^{56–58} Recently we developed a novel density-based energy decomposition analysis (DEDA) method⁵⁵ for intermolecular interactions within the framework of density functional theory (DFT). In comparison with other EDA approaches,^{18,21,22,48–54,59–63} which are all wave function based, this new DEDA method has the following two unique features: (1) It variationally calculates the total frozen density energy through constrained search; thus

Received: May 11, 2011

Published: October 18, 2011

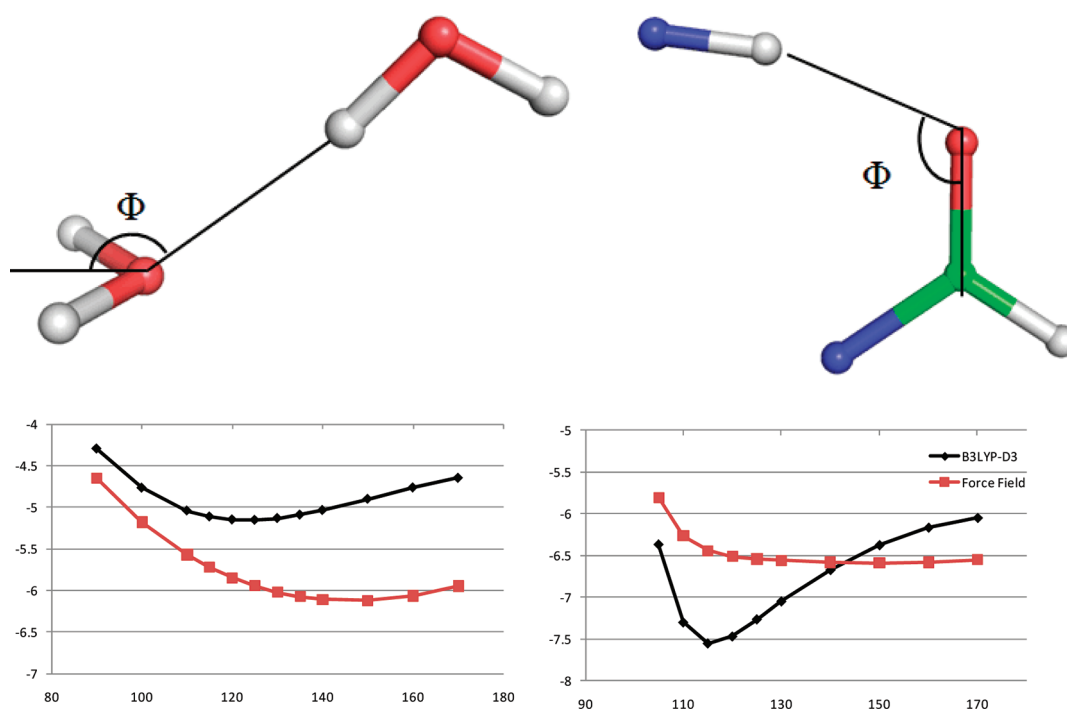


Figure 1. Illustration of the failure of classical atomic-charge-based force fields in modeling the hydrogen-bonding angular dependency. Please see ref 10 for a more detailed discussion of this problem. Upper panel: Φ angle for water and formamide dimers. Lower panel: binding energies along the Φ angle from high-level QM calculations (black) and force field calculations using atomic charge and LJ 12-6 model (red). Energy is in kcal/mol.

it enables a clean separation of the frozen density interaction energy from the density relaxation (i.e., polarization and charge transfer) contribution; and (2) the charge-transfer component is also calculated variationally based on the net electron flow in real space. These unique features enable the definition of each interaction component in DEDA to be more consistent with the typical physical picture employed in the classical force field description of intermolecular interactions and thus would make DEDA potentially more appropriate and helpful for force field development. By applying this method to the formamide dimer,⁵⁵ we showed earlier that the frozen density energy term is the dominant factor in determining the angular dependence of hydrogen bonding at the acceptor atom of the carbonyl group. This finding is quite different from the popular view regarding the origin of the hydrogen-bonding directionality.^{1,10–17} If it were generally true, it would mean that the difficulty for current nonpolarizable force fields to describe HB orientation is not due to the lack of explicit polarization or charge-transfer terms, which implicates that the problem of HB orientation specificity can be more approachable.

In this work, we have further improved the DEDA analysis protocol and made a systematic investigation on the directional dependence of HBs with both B3LYP-D3 and M06-2X-D3 functionals.^{64–70} The performance of both functionals to describe structures and binding interaction energies for a variety of hydrogen-bonding systems has been recently extensively tested and found to be excellent.^{70–74} Besides the formamide dimer, we have studied the hydrogen-bonding interactions between water and molecules containing sp nitrogen, aromatic oxygen and sp² and sp³ nitrogen, oxygen, and sulfur. The results clearly demonstrate that the frozen density interaction is the dominant factor in determining the HB orientation, while the sum of polarization and charge-transfer components shows very little HB directional

dependence. Then using the DEDA results as reference, we have examined the performance of several electrostatic and vdW interaction models and provided guidance for further force field development. Systematic parametrization protocols to determine parameters for more advanced electrostatic models based on QM calculations of monomers have been explored and presented, including the off-center charge model with electrostatic potential (ESP) fitting and distributed multipole model with distributed multipole analysis of Gaussian wavefunctions (GDMA) analysis,^{32,75} with or without taking account of charge penetration effects.^{76–79} For the vdW interaction, we have investigated the three most commonly employed potential energy functions,⁸⁰ i.e., LJ 12-6, buffered 14-7, and Buckingham potentials. In the following, we first give an introduction to the theory and methodology used in the current work and then present our results and discuss implications on further force field development.

2. THEORY AND METHODS

2.1. DEDA Method. Recently Wu et al.⁵⁵ developed a novel purely density-based EDA method for intermolecular binding within the framework of DFT. Here we first briefly review this scheme. To apply it to recent DFT methods, where the vdW attraction is taken into account to some extent either by the addition of empirical terms (i.e., the DFT-D methods)^{66,68,81,82} or the parametrization of the exchange–correlation functionals (such as X3LYP⁸³ and M06-2X functionals),⁶⁹ we lump the Pauli repulsion and the vdW attraction together to match the vdW interaction in classical force fields. Meanwhile, BSSE-corrected fragment energies and densities are employed to eliminate the basis set superposition error (BSSE) in the DEDA analysis.

Given the formation of the complex AB without distortion of the fragment geometry, as illustrated in Figure 2, the total

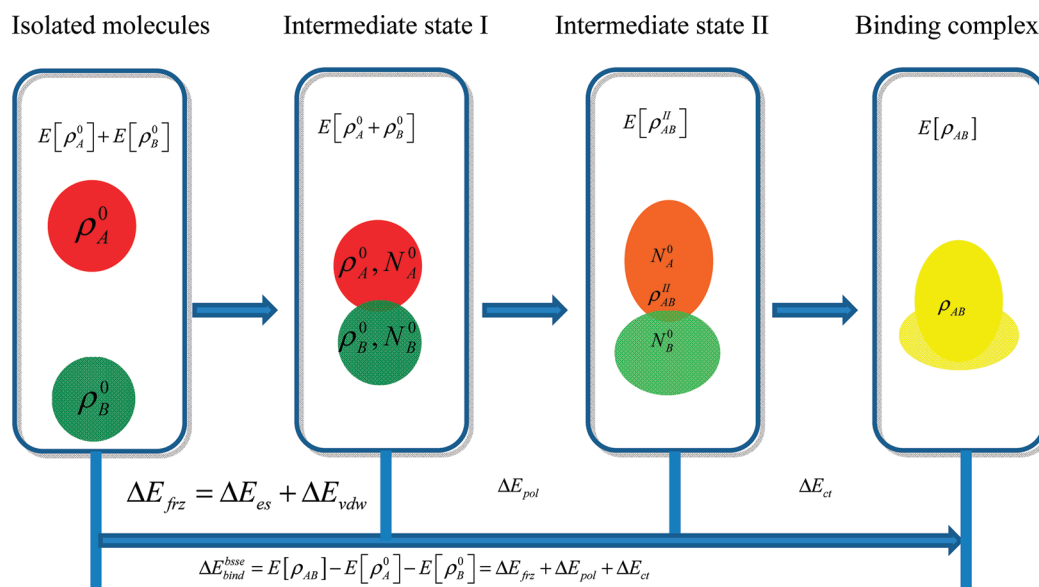
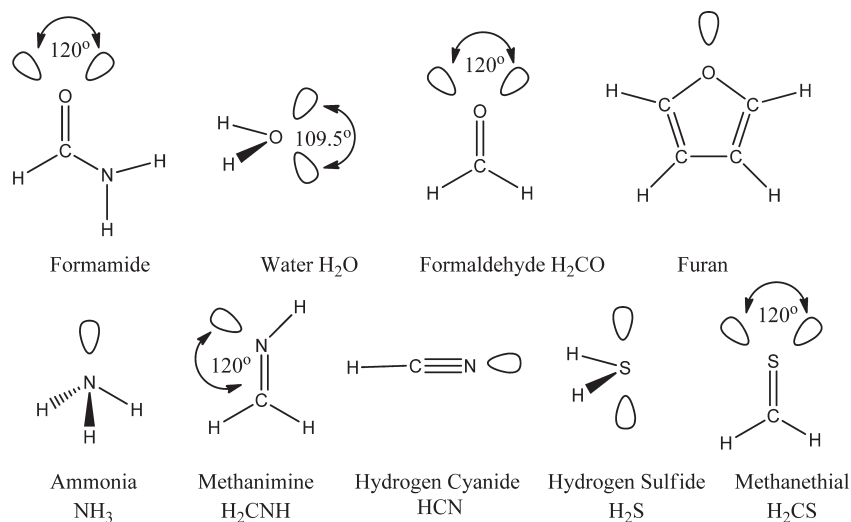


Figure 2. Illustration of the DEDA scheme.

Scheme 1. Illustration of the Placement of LP on Each Molecule for the off-Center Charge Model



BSSE-corrected binding energy can be decomposed into these contributions:

$$\Delta E_{\text{bind}}^{\text{BSSE}} = E[\rho_{\text{AB}}] - E[\rho_{\text{A}}^0] - E[\rho_{\text{B}}^0] = \Delta E_{\text{frz}} + \Delta E_{\text{pol}} + \Delta E_{\text{ct}} \quad (1)$$

where

$$\Delta E_{\text{frz}} = \Delta E_{\text{es}} + \Delta E_{\text{vdW}} = E[\rho_{\text{A}}^0 + \rho_{\text{B}}^0] - E[\rho_{\text{A}}^0] - E[\rho_{\text{B}}^0] \quad (2)$$

$$\Delta E_{\text{es}} = E_{\text{coulomb}}[\rho_{\text{A}}^0 + \rho_{\text{B}}^0] - E_{\text{coulomb}}[\rho_{\text{A}}^0] - E_{\text{coulomb}}[\rho_{\text{B}}^0] \quad (3)$$

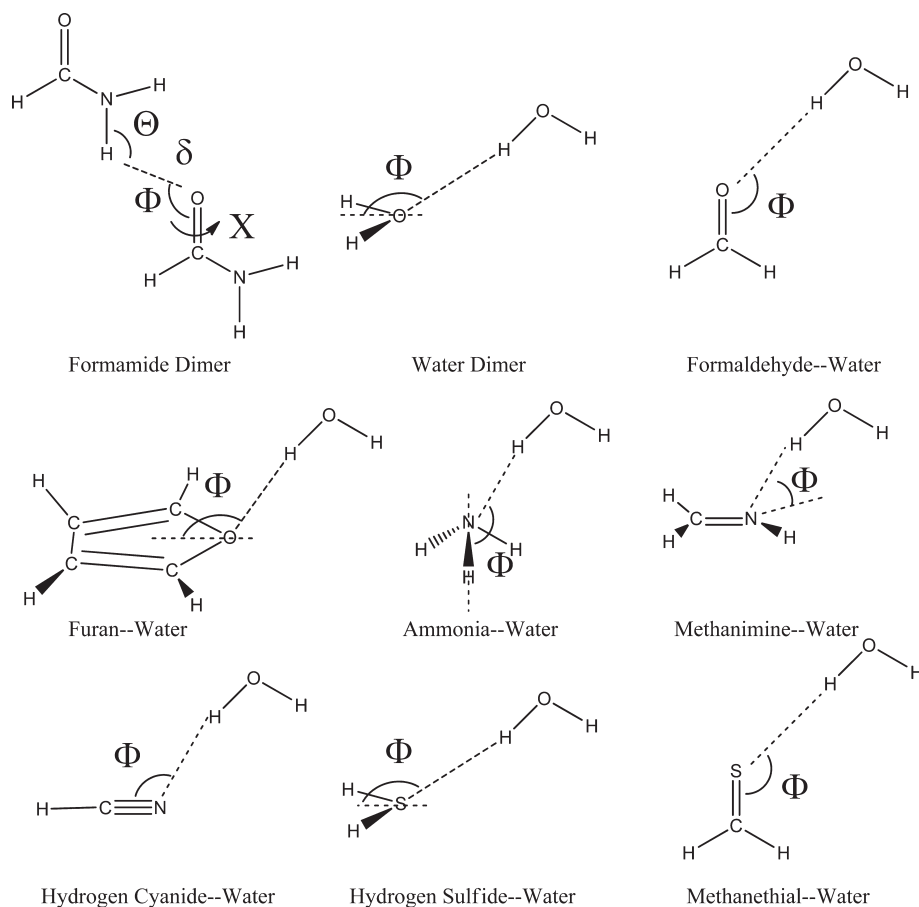
$$\Delta E_{\text{pol}} = E[\rho_{\text{AB}}^{\text{II}}] - E[\rho_{\text{A}}^0 + \rho_{\text{B}}^0] \quad (4)$$

$$\Delta E_{\text{ct}} = E[\rho_{\text{AB}}] - E[\rho_{\text{AB}}^{\text{II}}] \quad (5)$$

$E[\rho_{\text{AB}}]$ is the total energy of the binding complex AB, while $E[\rho_{\text{A}}^0]$ and $E[\rho_{\text{B}}^0]$ are total energies of noninteracting molecules A and B. They are determined with standard DFT calculations in which the energy and the density for each noninteracting fragment is calculated with all basis functions in the complex, as done in the standard counterpoise method.⁸⁴

As illustrated in Figure 2, the intermediate state I is the frozen density state, whose total density should be the superposition of two fragments' densities without any distortions. In DEDA,⁵⁵ the frozen density energy $E[\rho_{\text{A}}^0 + \rho_{\text{B}}^0]$ is variationally determined with a constrained search formalism,⁸⁵ i.e., $E[\rho_{\text{A}}^0 + \rho_{\text{B}}^0] = \min_{\rho} \rightarrow \rho_{\text{A}}^0 + \rho_{\text{B}}^0 E[\rho]$. It should be noted that in other EDA approaches,^{18,21,22,48–54,59–63} the Heitler–London (HL) antisymmetrization of two fragments' wave functions has been employed to represent such a frozen density state, in which the HL wave function is not variationally optimized and the

Scheme 2. HB Complexes and the Geometric HB Parameters (Denoted in Greek letters) Used For EDA and Model Evaluations



corresponding density does not correspond to the sum of fragments' densities.⁵⁴ Thus one most significant and novel feature of DEDA is to have a variationally determined frozen density energy through constrained search, which allows a clean separation of ΔE_{fz} from the density relaxation terms (ΔE_{pol} and ΔE_{ct}). As shown in eqs 2 and 3, ΔE_{fz} can be further decomposed into electrostatic and vdW interaction energies, ΔE_{es} and ΔE_{vdW} , which can be served as the reference to assess the performances of electrostatic and vdW interaction models, respectively. Dispersion interactions are not explicitly included in most DFT functionals, whose performance to describe midrange interactions can be strongly affected by the tail region of functionals⁸⁶ due to density overlap. Two recent developments to mitigate this deficiency within the GGA and hybrid functionals are: (1) adding an empirical term to model dispersion interactions (i.e., DFT-D method);^{66,81,82} and (2) taking the vdW interactions into account in the parametrization of the exchange–correlation functionals (such as for X3LYP⁸³ and M06-2X functionals).⁶⁹ For both DFT approaches, the resulting ΔE_{vdW} term in eq 2 would include both Pauli repulsion and vdW attraction contributions.

In order to determine ΔE_{pol} and ΔE_{ct} in eqs 4 and 5, the total energy for a second intermediate state $E[\rho_{\text{AB}}^{\text{II}}]$ is variationally calculated, in which the molecular density is relaxed without charge transfer using the constrained DFT method.⁸⁷ ΔE_{pol} accounts for the mutual polarization effect between the fragments, and ΔE_{ct} is the contribution to the total binding energy due to the charge transfer effect between the fragments. In DEDA,⁵⁵ these two

components are calculated based on density deformation and net electron flow in real space, respectively, and thus show a small basis set dependence. Furthermore, the above two unique features of DEDA⁵⁵ also make the resulting frozen energy and polarization components much less affected by the inherent delocalization error⁸⁸ (i.e., the enlarged self-interaction error for a fractional number of electrons)⁸⁹ inherent in most density functionals.

2.2. Electrostatic Models and Parameterization Protocols.

With the DEDA results as references, we have examined the performance of several electrostatic models, including point charge, off-center charge,^{25–30} and distributed multipole models.^{31–33} The ESP fitting method⁹⁰ has been employed to determine charge values for point charge and off-center charge models.

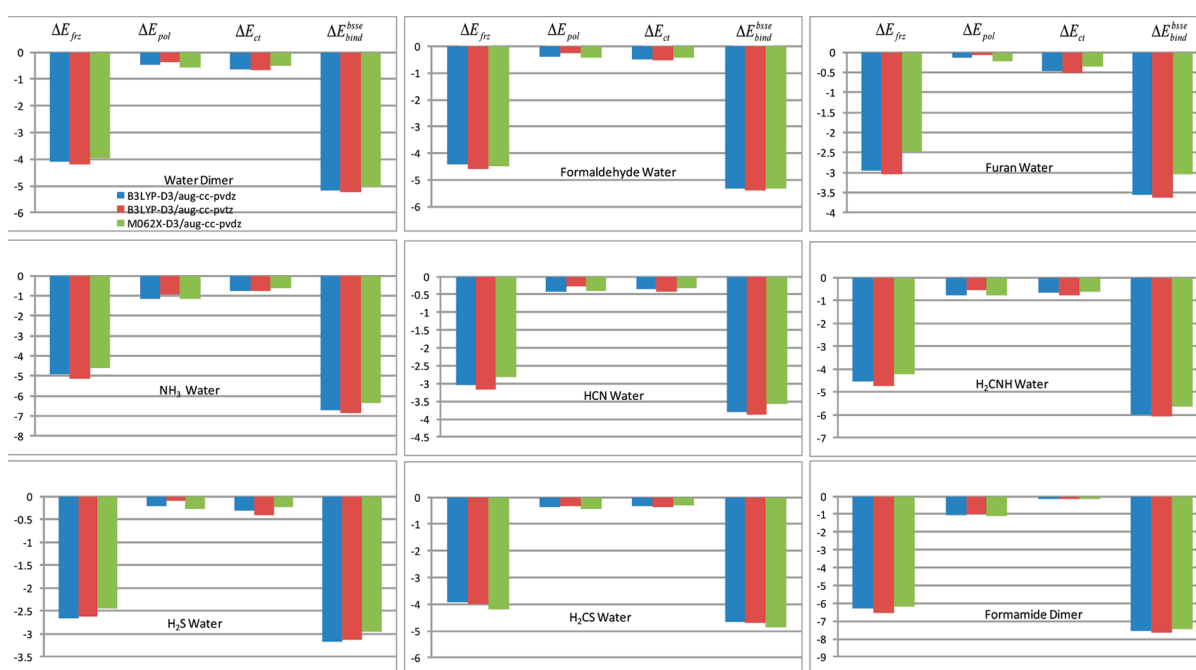
For the off-center charge model, extra point charges are placed on bond middle points and along the LP directions (see Scheme 1). We have found that adding extra charge sites at bond middle points improves the description of the directional dependence of hydrogen bonding. Determining the locations of LP charges is not trivial, and the optimal locations may vary with atom types.^{27,28} We have tried placing the LP charges either 0.35²⁶ or 0.74 Å³⁰ away from the corresponding nuclei and found that 0.74 Å in general leads to better agreement with the DEDA results and more meaningful point charge values. Therefore, only the results of 0.74 Å are presented.

To avoid unphysical charge values due to the ill-conditioning problem of the ESP charge fit, we employed the restrained

Table 1. Percentage of Energy Components and Total Binding Energy for Each HB Complex at Geometries Optimized at the B3LYP/aug-cc-pvdz Level^a

	B3LYP-D3/aug-cc-pvdz				B3LYP-D3/aug-cc-pvtz				M06-2X-D3/aug-cc-pvdz			
	ΔE_{frz}	ΔE_{pol}	ΔE_{ct}	ΔE_{bind}	ΔE_{frz}	ΔE_{pol}	ΔE_{ct}	ΔE_{bind}	ΔE_{frz}	ΔE_{pol}	ΔE_{ct}	ΔE_{bind}
water dimer	79	9	12	-5.16	80	7	13	-5.24	79	11	10	-5.02
H ₂ CO-H ₂ O	83	7	9	-5.31	85	5	10	-5.38	84	8	8	-5.33
furan-H ₂ O	83	4	13	-3.56	84	2	14	-3.63	81	7	12	-3.05
NH ₃ -H ₂ O	73	17	11	-6.73	75	14	11	-6.83	72	18	10	-6.37
HCN-H ₂ O	80	11	9	-3.80	82	7	11	-3.87	79	11	9	-3.57
H ₂ CNH-H ₂ O	76	13	11	-5.98	78	9	13	-6.07	75	14	11	-5.63
H ₂ S-H ₂ O	84	7	10	-3.16	84	3	13	-3.13	83	9	8	-2.95
H ₂ CS-H ₂ O	84	8	7	-4.67	85	7	8	-4.70	86	9	6	-4.86
formamide dimer	83	14	2	-7.55	85	13	2	-7.67	83	15	2	-7.46

^a ΔE_{frz} , ΔE_{pol} , and ΔE_{ct} are in percentage, and ΔE_{bind} is in kcal/mol.

**Figure 3.** Total binding energy and its components for each HB complex at geometries optimized at the B3LYP/aug-cc-pvdz level.

electrostatic potential (RESP) fit by adding a hyperbolic penalty function to the least-squares sum:⁹¹

$$\begin{aligned} \chi^2 &= \chi_{\text{esp}}^2 + \chi_{\text{rest}}^2 \\ &= \sum_i (V_i - \hat{V}_i)^2 + a \sum_j [(q_j^2 + b^2)^{1/2} - b], \end{aligned} \quad (6)$$

where V_i and \hat{V}_i are the electrostatic potentials calculated by QM and the off-center atomic model at grid i , and q_j is the point charge at charge site j . Parameters $a = 0.005$ au and $b = 0.1$ define the strength of the restraint and the tightness of the hyperbola around the minimum, respectively.⁹¹

For distributed multipole models,^{31–33} we have tested multipole moments up to quadrupole, and the results confirm that quadrupole is necessary to yield an accurate description of the intermolecular electrostatics for HB complexes. In comparison with the distributed multipole model employed in the AMOEBA

force field,^{4,33,92} we also added distributed multipoles at bond middle points, which slightly improves the results. The distributed multipoles were calculated with the GDMA program version 2.2,⁷⁵ using the formatted checkpoint file produced by Gaussian03⁹³ as input. All distributed multipole analyses, except for the furan molecule, were performed with the analytical DMA algorithm by setting keywords “SWITCH 0” and “Radius H 0.35”. For the furan molecule, the analytical DMA algorithm breaks down due to a known stability problem with large basis sets.⁷⁵ Therefore, numerical DMA was performed for the furan molecule with the default setting in the program. In agreement with a previous study,⁷⁸ we find that the analytical DMA algorithm in general yields more accurate results on electrostatic potentials and intermolecular electrostatic interactions than numerical DMA.

Charge penetration^{76–79} has been known to make significant contribution to intermolecular electrostatic interaction in the short range. In order to examine its effect on the description of

the directional dependence of hydrogen bonding, we have replaced each point charge with a smeared charge, which consists of a nuclear charge Z and an exponential charge density $\rho(r) = (qa^3/8\pi)e^{-ar}$, in which r is the distance to the charge center and a defines the width of the charge distribution. Accordingly, the electrostatic interaction between two smeared charges can be

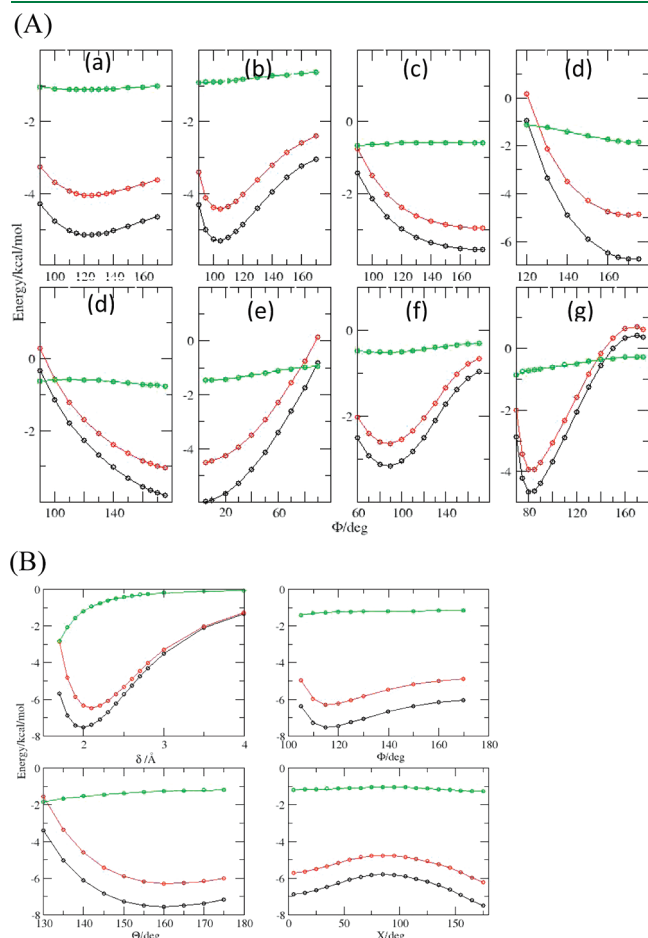


Figure 4. Total binding energy and its components from B3LYP-D3/aug-cc-pvdz DEDA along the HB angles (see Scheme 2). Black: total binding energy; red: frozen density energy; and green: electronic relaxation energy (polarization and charge transfer). Panel A: (a) Water dimer; (b) formaldehyde–water; (c) furan–water; (d) ammonia–water; (e) hydrogen cyanide–water; (f) methanimine–water; and (g) hydrogen sulfide–water. Panel B: formamide dimer.

written as:

$$E^{\text{chg} - \text{chg}} = \frac{q_A q_B}{R} [1 - f(a, b, R) - f(b, a, R)] + \frac{q_A Z_B}{R} g(a, R) + \frac{q_B Z_A}{R} g(b, R) + \frac{Z_A Z_B}{R} \quad (7)$$

where

$$f(a, b, R) = e^{-aR} \frac{b^4}{(b^2 - a^2)^2} \left(1 - \frac{2a^2}{b^2 - a^2} + \frac{aR}{2} \right)$$

and

$$g(a, R) = 1 - e^{-aR} \left(1 + \frac{aR}{2} \right)$$

The interaction between one smeared charge and one point dipole is

$$E^{\text{chg} - \text{dip}} = -(Z_A + \lambda_3 q_A) \vec{\mu}_B \cdot \frac{\vec{R}}{R^3} \quad (8)$$

where $\lambda_3 = 1 - e^{-aR} - aR e^{-aR} - (a^2 R^2/2) e^{-aR}$ and the interaction between one smeared charge and one traceless point quadrupole is

$$E^{\text{chg} - \text{quadrupole}} = (Z_A + \lambda_5 q_A) \sum \Theta_{\alpha\beta}^B \frac{R_\alpha R_\beta}{R^5} \quad (9)$$

where $\lambda_5 = \lambda_3 - 1/6 a^3 R^3 e^{-aR}$, and $\Theta_{\alpha\beta}^B$ is the traceless quadrupole moment at site B.

For each atomic center, the nuclear charge Z_A is taken as the number of valence electrons, for example $Z = 4$, for the carbon atom. Parameter a for each charge site is determined by minimizing the electrostatic potential difference between quantum mechanical calculations and the damped multipolar expansion over a set of grid points.⁷⁸ The minimization is performed with the modified “potential” subprogram in TINKER 5.0.⁹⁴ The fitting grid consists of 10 layers with a spacing of 0.25 Å. The first layer starts at the distance of half of the vdW radii⁹⁵ to the atomic center.

2.3. vdW Interaction Models. For vdW interactions, we have investigated the three most commonly employed potential energy functions,⁸⁰ i.e. LJ 12-6, Buckingham, and buffered 14-7 potentials:

$$E_{ij}^{\text{LJ}} = \varepsilon_{ij} \left(\frac{R_{\min, ij}^{12}}{R_{ij}^{12}} - \frac{2R_{\min, ij}^6}{R_{ij}^6} \right) \quad (10)$$

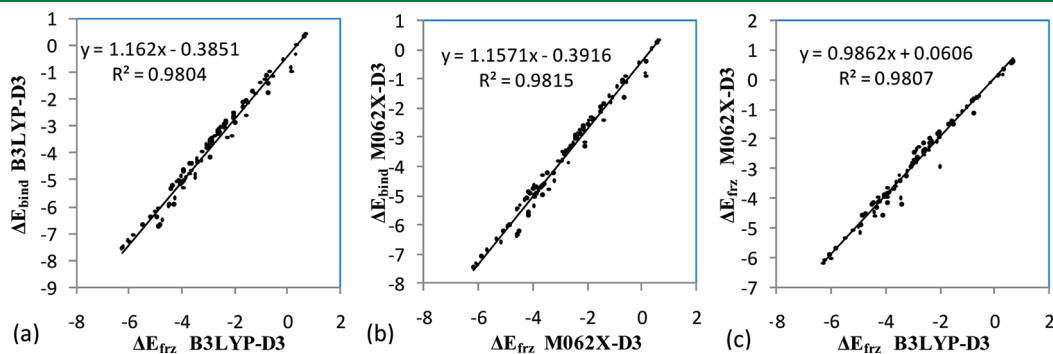


Figure 5. Linear regressions for total binding energy and frozen density interaction results (in kcal/mol) along HB angles: (a) For B3LYP-D3, ΔE_{bind} vs ΔE_{frz} ; (b) for M06-2X-D3, ΔE_{frz} vs ΔE_{bind} ; and (c) for ΔE_{frz} B3LYP-D3 vs M06-2X-D3.

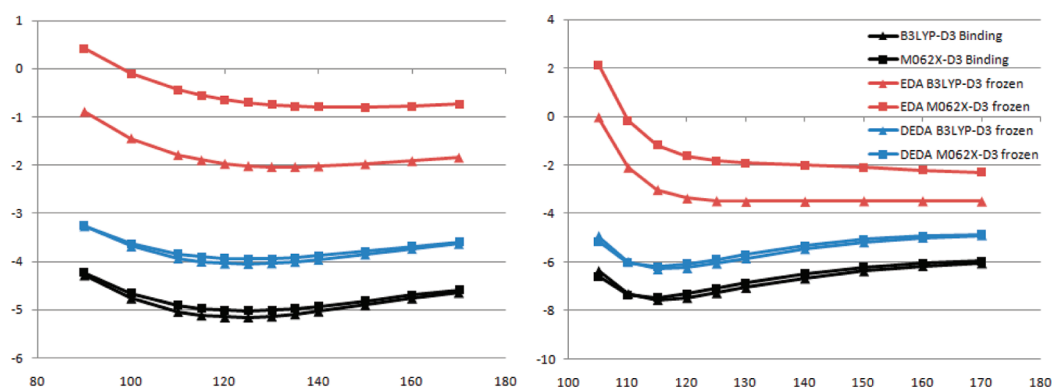


Figure 6. Comparison of frozen energies in kcal/mol along the Φ angle (see Figure 1) for water (left) and formamide (right) dimers with DEDA and the wave function-based EDA. The latter employs the HL antisymmetrization of two fragments' wave functions to represent the frozen density state.

$$E_{ij}^{\text{Buck}} = \varepsilon_{ij} \left(A e^{-BR_{ij}/R_{\min,ij}} - C \frac{R_{\min,ij}^6}{R_{ij}^6} \right) \quad (11)$$

and

$$E_{ij}^{\text{buff}} = \varepsilon_{ij} \left(\frac{1 + \delta}{\rho_{ij} + \delta} \right)^7 \left(\frac{1 + \gamma}{\rho_{ij}^7 + \gamma} - 2 \right) \quad (12)$$

In the above equations, R_{ij} is the distance between atoms i and j , ε_{ij} is the potential well depth, $R_{\min,ij}$ is the minimum energy distance, and $\rho_{ij} = R_{ij}/R_{\min,ij}$ in eq 12. The combination rules for eqs 10 and 11 are

$$R_{\min,ij} = r_i^0 + r_j^0 \text{ and } \varepsilon_{ij} = \sqrt{\varepsilon_i^0 \varepsilon_j^0} \quad (13)$$

For eq 12, it is

$$R_{\min,ij} = 2 \frac{(r_i^0)^3 + (r_j^0)^3}{(r_i^0)^2 + (r_j^0)^2} \text{ and } \varepsilon_{ij} = \frac{4\varepsilon_i^0 \varepsilon_j^0}{((\varepsilon_i^0)^{1/2} + (\varepsilon_j^0)^{1/2})^2} \quad (14)$$

where r_i^0 and ε_i^0 are vdW parameters for the atom i .

While eq 10 (LJ 12-6) is the most popular functional form for vdW interactions, and eq 11 (Buckingham) and eq 12 (buffered 14-7) are used in MM3²⁴ and AMOEBA³³ force fields, respectively. Following the referenced force fields,^{24,33} we took $A = 18400.0$, $B = 12$, and $C = 2.25$ in eq 11, and $\delta = 0.07$ and $\gamma = 0.12$ in eq 12; r_i^0 and ε_i^0 depend on functional forms and were taken directly from the corresponding force field (OPLS-AA for LJ 12-6, MM3 for Buckingham, and AMOEBA for buffered 14-7).

3. COMPUTATIONAL DETAILS

3.1. HB Complexes. The HB complexes investigated in this work are illustrated in Scheme 2. Except for the formamide dimer, which was shown to be representative of HBs found in protein side chains and main chains,¹⁰ all other complexes have water as the HB donor. The HB acceptors, as shown in Scheme 1, represent a set of molecules containing sp nitrogen and aromatic oxygen and sp² and sp³ nitrogen, oxygen, and sulfur. With the internal coordinates of monomers frozen and other HB geometric parameters held fixed at the optimized dimer values, we scanned along the chosen HB geometric parameters (denoted by

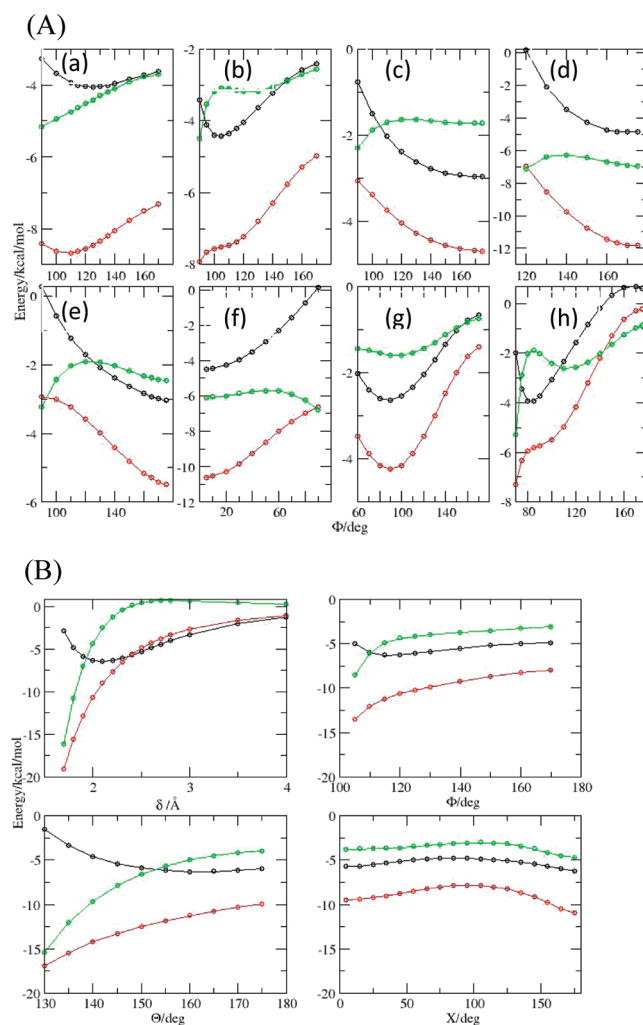


Figure 7. Frozen density energy, electrostatics, and vdW energy (in reversed sign) from B3LYP-D3 EDA along the HB angles (see Scheme 2). Black: frozen density energy; red: electrostatics; and green: vdW energy in reversed sign. Other descriptions are the same as those in Figure 4.

Greek letters in Scheme 2) to generate hydrogen-bonded dimers for the DEDA calculations and model evaluations. The internal geometry of each fragment was optimized at the B3LYP/aug-cc-pvdz level with Gaussian 03.⁹³ We do not consider energy

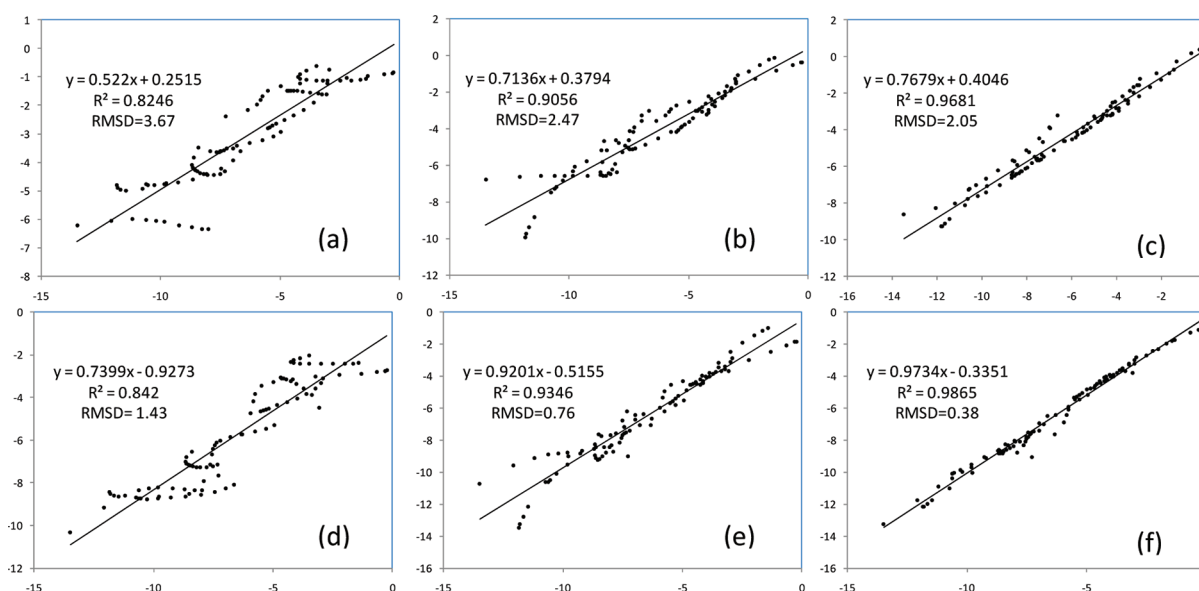


Figure 8. Results of linear regressions of ΔE_{es} (in kcal/mol) between each electrostatic model and B3LYP-D3 DEDA results in describing HB directional dependence: (a) Atomic charge model; (b) off-center point charge model; (c) distributed multipole model; (d) atomic smeared charge model; (e) off-center smeared charge model; and (f) distributed multipole model with smeared charge. The x-axes correspond to the results from B3LYP-D3/aug-ccpvdz DEDA.

contributions from internal geometry distortions, i.e., the preparation energy.

3.2. DFT Calculations and EDA Analyses. DEDA were carried out at the B3LYP^{65,67,81}-D3/aug-cc-pvdz level and M06-2X⁶⁹-D3/aug-cc-pvdz level with a development version of NWCHEM.⁹⁶ The notation D3 indicates that the total binding energy and the vdW component from DEDA are corrected for each DFT functional by adding the dispersion energy calculated from the newest version of Grimme's method.⁶⁶ While the correction is significant for B3LYP, it is in general less than 0.1 kcal/mol for M06-2X. For the M06-2X calculations, the fine grid as implemented in NWCHEM is applied to reduce the integration grid errors when using meta-generalized-gradient approximation functionals.^{97–99} The Beck's integration scheme¹⁰⁰ is used for the weighting function for charge partition.⁵⁵ To test the influence of the basis sets, calculations at the B3LYP-D3/aug-cc-pvtz level were also performed.

4. RESULTS AND DISCUSSION

In this section we will first present the DEDA results for various HB complexes and our finding that frozen density interaction is the key factor in determining angular dependence of HBs. Then by using the B3LYP DEDA results as reference, we will evaluate various electrostatic and vdW models in their descriptions of the directional hydrogen bonding.

4.1. DEDA of Hydrogen-Bonding Directional Dependence. As illustrated in Figure 2, the total hydrogen-bonding interaction energy can be cleanly divided into four physically meaningful components, i.e., electrostatics, vdW, polarization, and charge transfer. The sum of the electrostatics and vdW terms constitutes the frozen density interaction energy term, which in principle has already been taken into account in nonpolarizable force fields; on the other hand, the polarization and charge-transfer terms come from the electronic relaxation, whose explicit description by force fields needs polarizable models and other advanced treatment.

The DEDA results of the nine hydrogen-bonding complexes (Scheme 2) at their optimal HB configurations are presented in Table 1 and Figure 3. We can see that among different functionals and basis sets, including B3LYP-D3/aug-cc-pvdz, B3LYP-D3/aug-cc-pvtz, and M06-2X-D3/aug-cc-pvdz, the results are very consistent. At the optimal HB configuration, the electronic relaxation energy is about 20% of the total binding energy, which is certainly not negligible. Among the nine complexes, the ammonia–water complex has the largest contribution from electronic relaxation energy, which accounts for about 28% of the total HB binding energy. As to the influence of the basis set, the results from aug-cc-pvdz and aug-cc-pvtz are in close agreement with each other, indicating that DEDA is not sensitive to the size of the basis set. Binding energies from M06-2X-D3 are also in general very consistent with those from B3LYP-D3. We note that without the dispersion correction, an important ingredient of HB interaction,⁷¹ B3LYP binding energies are about 1–2 kcal/mol weaker than those from M06-2X. The results in Table 1 and Figure 3 further confirm that the frozen density interaction plays a major role in determining the hydrogen-bonding strength.

The DEDA results of the nine hydrogen-bonding complexes with respect to their respective hydrogen-bonding angles defined in Scheme 2 are presented in Figures 4 and 5. For all nine complexes, we can clearly see that the frozen density interaction energy (red in Figure 4) shares the same trend as the total binding energy (black), while the electronic relaxation energy (green) is quite flat along those defined angles. In fact, as shown in Figure 5a, we see a strong correlation between ΔE_{frz} and ΔE_{bind} with respect to Φ . Moreover, the resemblance between ΔE_{frz} and ΔE_{bind} is not limited to the B3LYP-D3 functional. In Figure 5b we can see there is also a strong correlation between ΔE_{frz} and ΔE_{bind} for M06-2X-D3. The consistency of ΔE_{frz} between two different functionals is demonstrated in Figure 5c, in which the linear regression yields a close-to-unity slope with R^2 nearly one and residual close to zero. Thus our DEDA results clearly show that the frozen density interaction term is the dominant factor in

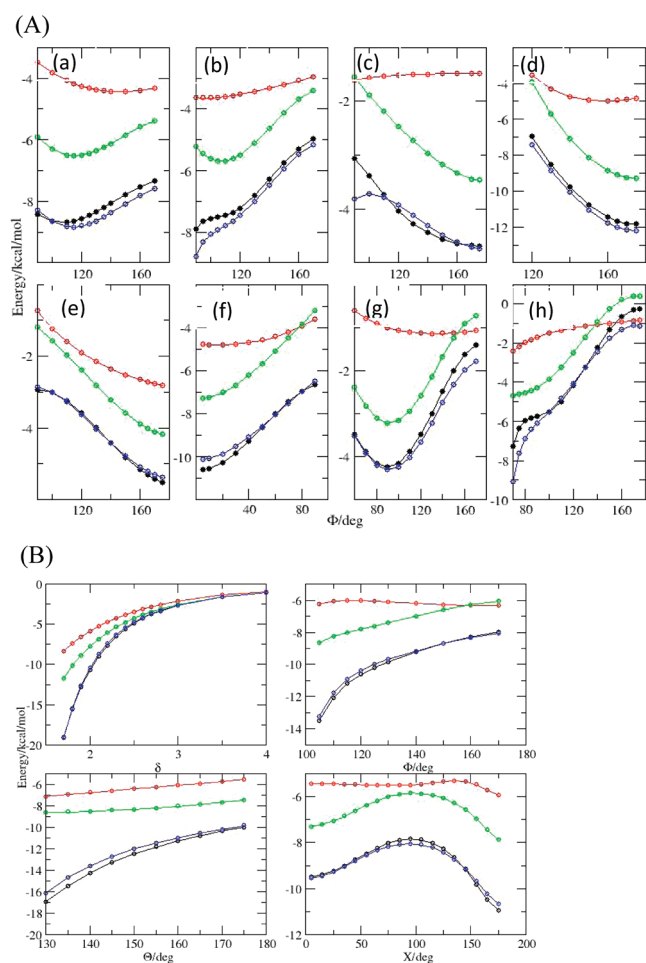


Figure 9. ΔE_{es} (in kcal/mol) along the HB angles obtained from B3LYP-D3 DEDA and different electrostatic models. Black circle: DEDA; red: atomic charge model; green: distributed multipole model; and blue: distributed multipole model with smeared charge to take account of charge penetration effects. Other descriptions are the same as those in Figure 4.

determining the hydrogen-bonding orientation, while the sum of polarization and charge-transfer terms shows very little HB directional dependence. It should be noted that this finding is quite different from the current dominant view regarding the origin of hydrogen-bonding directionality.^{1,10–17}

In order to find out whether similar results can be observed with wave function-based EDA methods, we also performed the same analysis with the wave function based EDA, in which the HL antisymmetrization of two fragments' wave functions has been employed to represent the frozen density state. The comparison of results between EDA and DEDA on the water and formamide dimers is presented in Figure 6, and it shows three clear distinctions: (1) ΔE_{frz} from EDA is significantly smaller, which implies the contribution from electronic relaxation energy to ΔE_{bind} is significantly larger for EDA; (2) there is no strong correlation between ΔE_{frz} and ΔE_{bind} for EDA, which suggests that polarization and charge transfer play important roles in HB orientation based on wave function-based EDA analysis; and (3) EDA has a much larger difference of ΔE_{frz} between different DFT functionals than DEDA. These distinctions clearly demonstrate important novel features of the DEDA approach and may lend

DEDA some unique advantages for force field development over the wave function-based EDA methods. Specifically, our DEDA results indicate that the problem of describing HB orientation in current nonpolarizable force fields is not due to the lack of explicit polarization or charge-transfer terms but comes from deficiencies in electrostatic and vdW models. As shown in Figure 7, both electrostatics and vdW interactions are important in determining the HB directionality, and the optimal binding angle results from a balance between them. In next two subsections, using the DEDA results as reference, we will examine several electrostatic and vdW models in order to provide some guidance for further force field development.

4.2. Examination of Electrostatic Models. To go beyond the atomic charge model for better describing electrostatics interactions, there have been substantial efforts on developing off-center charge models,^{25–30} which place extra point charges at sites other than the nuclei and models employing higher multipole moments.^{31–33,101}

The details of implementation and parametrization can be quite diverse among different approaches. Here our strategy is to avoid the employment of dimerization data in parametrization; instead we derive all parameters by only using electrostatic properties of monomers. Thus our parametrization procedure can be directly applied to any molecules, and the results will not be biased toward the HB complexes investigated here. As described in Section 2, we use RESP fitting to derive charges for the off-center charge model and the GDMA method⁷⁵ for distributed multipoles up to quadrupole.

We have examined the performance of various electrostatic models to describe directional hydrogen bonding. Correlations between ΔE_{es} from electrostatic models and the corresponding ΔE_{es} from B3LYP-D3/aug-cc-pvdz DEDA are shown in Figure 8. For the atomic charge model (Figure 8a), a slope of ~ 0.5 indicates a significant underestimation of electrostatic interactions in comparison with the DEDA results, and the root-mean-square deviation (RMSD) is quite large (3.7 kcal/mol). A more serious problem for the atomic charge model is that the electrostatic interaction curves are too flat for most of the HB complexes as shown in Figure 9, which is directly responsible for its problem of describing HB directionality. From Figure 8b and c, we can see that the off-center charge model and the distributed multipole model make significant improvement over the atomic charge model: Slopes are larger, RMSDs are smaller, and R^2 values are significantly close to unity. As shown in Figure 9, ΔE_{es} from the distributed multipole model has a good agreement with the DEDA results, except for a systematic shift that is largely due to the charge penetration effect.

The influence of charge penetration on the short-range intermolecular electrostatic interactions is well-known³² and has recently attracted much attention.^{56,76–79,102,103} To account for the charge penetration effect, we replace each point charge with a smeared charge, which consists of a negative exponential charge density and a positive point charge at the atomic center.⁵⁶ The net effect of smeared charges increases the strength of electrostatic interactions at the short-range. Here we have only considered the charge penetration effects for monopoles because it was shown that introducing higher order damping only yields minor improvement.⁷⁸ As demonstrated in Figure 8, with the smeared charge model, the electrostatic interactions become much stronger, and the RMSDs are all greatly reduced for all three models. There is an excellent agreement between the DEDA results and those from the smeared charge distributed multipole model, as indicated by the close-to-unity slope and the

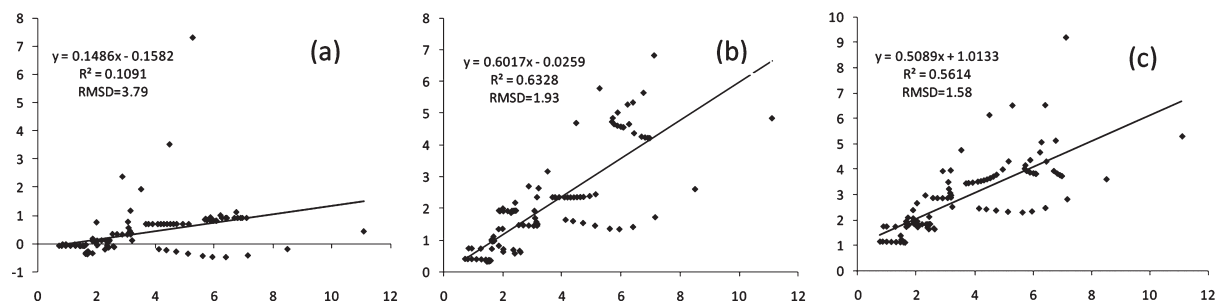


Figure 10. Linear regression results for ΔE_{vdW} (in kcal/mol) between each vdW model and B3LYP-D3 DEDA. (a) LJ 12-6; (b) buffered 14-7; and (c) Buckingham. The x -axes correspond to the results from B3LYP-D3 DEDA.

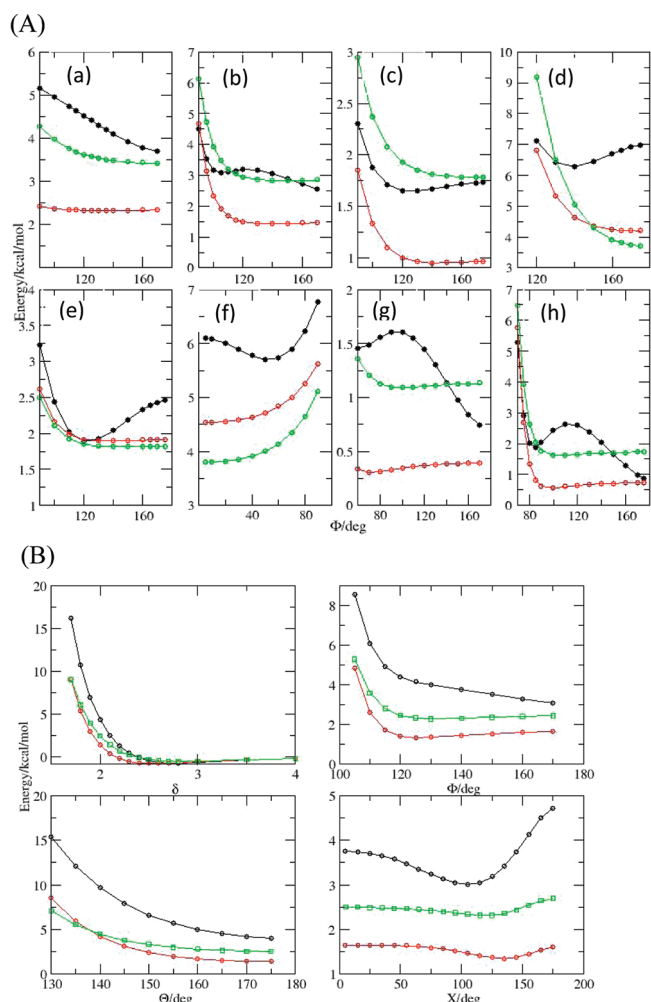


Figure 11. vdW interactions along the HB angles obtained from B3LYP-D3 DEDA and various vdW models. Black: DEDA; red: buffered 14-7; and green: Buckingham. Other descriptions are the same as those in Figure 4.

small RMSD in Figure 8f. Meanwhile, by comparing the values of R^2 in Figure 8a and d, we can see that incorporating the charge penetration effects alone would not be sufficient to relieve the problem of describing HB orientation in the atomic charge model. Among all models examined, Figures 8 and 9 clearly show that the smeared charge distributed multipole model (up to quadrupole), which takes account of charge

penetration effect, has the best agreement with the corresponding DEDA results.

4.3. Examination of vdW Models. As shown in Figure 7, not only electrostatics but also vdW interactions can be important in determining the HB directionality. Here we have investigated three commonly employed vdW force field functions. The vdW parameters were directly taken from OPLS-AA,⁷ AMOEBA,⁹² and MM3²⁴ force fields, respectively, to test the LJ12-6, buffered 14-7, and Buckingham potentials. These parameters are in general obtained by fitting to experimentally measured properties or high-level ab initio QM calculations of dimerization energies. Thus we do not expect that the vdW function forms with parameters taken directly from those force fields have a good agreement with the DEDA results. Nevertheless, such a comparison may shed light on the appropriateness of vdW function forms to describe the angular dependence of hydrogen bonding.

The correlations between ΔE_{vdW} from vdW models and those from B3LYP-D3 DEDA are presented in Figure 10. Among the three models, LJ 12-6 model is clearly the worst and does not correlate at all with the corresponding DEDA results. The performances of buffered 14-7 and Buckingham potentials are comparable. Both models have some weak correlations with the DEDA results but are still not satisfactory in describing the angular dependency of vdW interaction, as shown in Figure 11. To significantly improve the description of vdW interactions, a systematic parametrization protocol should be developed, and more advanced vdW models may also need to be explored. We are currently carrying out studies along this direction and will present our results in a future account.

5. CONCLUSIONS

In this work, we have further improved the density-based energy decomposition analysis (DEDA) method and made a systematic investigation about the directional dependence hydrogen bonding with both B3LYP-D3 and M06-2X-D3 functionals. Our results clearly demonstrate that frozen density interaction energy term is the key factor in determining the HB orientation, while the density relaxation energy, including both polarization and charge-transfer components, shows very little HB directional dependence. This finding is quite different from the current dominant view regarding the origin of hydrogen-bonding directionality and would not be obtained with wave function-based EDA approaches. Using the DEDA results as reference, we have examined several electrostatic and vdW interaction models and demonstrated that the main deficiency coming from the atomic point charge model can be overcome largely

by the introduction of extra charge sites or higher order multipole moments. Among all electrostatic models explored, the smeared charge distributed multipole model (up to quadrupole) is found to have the best agreement with the corresponding DEDA results. Systematic parametrization protocols to determine parameters for more advanced electrostatic models based on QM calculations of monomers have been presented. We have also shown that vdWs models still need further improvement to better model the directional dependence of hydrogen bonding.

AUTHOR INFORMATION

Corresponding Author

*E-mail: yingkai.zhang@nyu.edu.

ACKNOWLEDGMENT

This work carried out in part at NYU was supported by NIH (R01-GM079223) and NSF-MRSEC (DMR-0820341). Research carried out in part at the Center for Functional Nanomaterials was supported by the U.S. Department of Energy, Office of Basic Energy Sciences under contract no. DE-AC02-98CH10886. We thank NYU-ITS and CFN for providing computational resources.

REFERENCES

- Freddolino, P. L.; Harrison, C. B.; Liu, Y. X.; Schulten, K. *Nat. Phys.* **2010**, *6*, 751–758.
- Jorgensen, W. L. *Science* **2004**, *303*, 1813–1818.
- Karplus, M.; McCammon, J. A. *Nat. Struct. Biol.* **2002**, *9*, 646–652.
- Ponder, J. W.; Case, D. A. *Adv. Protein Chem.* **2003**, *66*, 27–+.
- van Gunsteren, W. F.; Bakowies, D.; Baron, R.; Chandrasekhar, I.; Christen, M.; Daura, X.; Gee, P.; Geerke, D. P.; Glatli, A.; Hunenberger, P. H.; Kastenholz, M. A.; Ostenbrink, C.; Schenk, M.; Trzesniak, D.; van der Vegt, N. F. A.; Yu, H. B. *Angew. Chem., Int. Ed.* **2006**, *45*, 4064–4092.
- Mackereell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M. *J. Phys. Chem. B* **1998**, *102*, 3586–3616.
- Jorgensen, W. L.; Maxwell, D. S.; TiradoRives, J. *J. Am. Chem. Soc.* **1996**, *118*, 11225–11236.
- Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. *J. Am. Chem. Soc.* **1995**, *117*, 5179–5197.
- Wang, J. M.; Cieplak, P.; Kollman, P. A. *J. Comput. Chem.* **2000**, *21*, 1049–1074.
- Morozov, A. V.; Kortemme, T.; Tsemekhman, K.; Baker, D. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 6946–6951.
- Best, R. B.; Hummer, G. *J. Phys. Chem. B* **2009**, *113*, 9004–9015.
- Kortemme, T.; Morozov, A. V.; Baker, D. *J. Mol. Biol.* **2003**, *326*, 1239–1259.
- Choi, H.; Kang, H.; Park, H. *J. Phys. Chem. B* **2010**, *114*, 2980–2987.
- Day, R.; Paschek, D.; Garcia, A. E. *Proteins: Struct., Funct., Bioinf.* **2010**, *78*, 1889–1899.
- Cao, Z. X.; Lin, Z. X.; Wang, J.; Liu, H. Y. *J. Comput. Chem.* **2009**, *30*, 645–660.
- Lii, J. H.; Allinger, N. L. *J. Phys. Chem. A* **2008**, *112*, 11903–11913.
- Freddolino, P. L.; Park, S.; Roux, B.; Schulten, K. *Biophys. J.* **2009**, *96*, 3772–3780.
- Umeyama, H.; Morokuma, K. *J. Am. Chem. Soc.* **1977**, *99*, 1316–1332.
- Grishaev, A.; Bax, A. *J. Am. Chem. Soc.* **2004**, *126*, 7281–7292.
- Bissantz, C.; Kuhn, B.; Stahl, M. *J. Med. Chem.* **2010**, *53*, 5061–5084.
- Khaliullin, R. Z.; Cobar, E. A.; Lochan, R. C.; Bell, A. T.; Head-Gordon, M. *J. Phys. Chem. A* **2007**, *111*, 8753–8765.
- Chen, W.; Gordon, M. S. *J. Phys. Chem.* **1996**, *100*, 14316–14328.
- Choi, H.; Kang, H.; Park, H. *J. Comput. Chem.* **2010**, *31*, 897–903.
- Allinger, N. L.; Yuh, Y. H.; Lii, J. H. *J. Am. Chem. Soc.* **1989**, *111*, 8551–8566.
- Cieplak, P.; Caldwell, J.; Kollman, P. *J. Comput. Chem.* **2001**, *22*, 1048–1057.
- Dixon, R. W.; Kollman, P. A. *J. Comput. Chem.* **1997**, *18*, 1632–1646.
- Karamertzanis, P. G.; Pantelides, C. C. *Mol. Simul.* **2004**, *30*, 413–436.
- Tschampel, S. M.; Kennerty, M. R.; Woods, R. J. *J. Chem. Theory Comput.* **2007**, *3*, 1721–1733.
- Wang, Z. X.; Zhang, W.; Wu, C.; Lei, H. X.; Cieplak, P.; Duan, Y. *J. Comput. Chem.* **2006**, *27*, 781–790.
- Zhao, D. X.; Liu, C.; Wang, F. F.; Yu, C. Y.; Gong, L. D.; Liu, S. B.; Yang, Z. Z. *J. Chem. Theory Comput.* **2010**, *6*, 795–804.
- Buckingham, A. D.; Fowler, P. W. *Can. J. Chem.* **1985**, *63*, 2018–2025.
- Stone, A. J. *The theory of intermolecular forces*; Oxford University Press: Oxford, U.K., 1996.
- Ren, P. Y.; Ponder, J. W. *J. Phys. Chem. B* **2003**, *107*, 5933–5947.
- Warshel, A.; Levitt, M. *J. Mol. Biol.* **1976**, *103*, 227–249.
- Ren, P. Y.; Ponder, J. W. *J. Comput. Chem.* **2002**, *23*, 1497–1506.
- Xie, W. S.; Pu, J. Z.; MacKerell, A. D.; Gao, J. L. *J. Chem. Theory Comput.* **2007**, *3*, 1878–1889.
- Thole, B. T. *Chem. Phys.* **1981**, *59*, 341–350.
- Sprink, M.; Klein, M. L. *J. Chem. Phys.* **1988**, *89*, 7556–7560.
- Straatsma, T. P.; Mccammon, J. A. *Chem. Phys. Lett.* **1990**, *167*, 252–254.
- van Maaren, P. J.; van der Spoel, D. *J. Phys. Chem. B* **2001**, *105*, 2618–2626.
- Lamoureux, G.; Roux, B. *J. Chem. Phys.* **2003**, *119*, 3025–3039.
- Yu, H. B.; Hansson, T.; van Gunsteren, W. F. *J. Chem. Phys.* **2003**, *118*, 221–234.
- Anisimov, V. M.; Lamoureux, G.; Vorobyov, I. V.; Huang, N.; Roux, B.; MacKerell, A. D. *J. Chem. Theory Comput.* **2005**, *1*, 153–168.
- Rappe, A. K.; Goddard, W. A. *J. Phys. Chem.* **1991**, *95*, 3358–3363.
- Rick, S. W.; Stuart, S. J.; Berne, B. J. *J. Chem. Phys.* **1994**, *101*, 6141–6156.
- York, D. M.; Yang, W. T. *J. Chem. Phys.* **1996**, *104*, 159–172.
- Patel, S.; Mackerell, A. D.; Brooks, C. L. *J. Comput. Chem.* **2004**, *25*, 1504–1514.
- Morokuma, K. *J. Chem. Phys.* **1971**, *55*, 1236–&.
- Kitaura, K.; Morokuma, K. *Int. J. Quantum Chem.* **1976**, *10*, 325–340.
- Stevens, W. J.; Fink, W. H. *Chem. Phys. Lett.* **1987**, *139*, 15–22.
- Frey, R. F.; Davidson, E. R. *J. Chem. Phys.* **1989**, *90*, 5555–5562.
- Bagus, P. S.; Illas, F. *J. Chem. Phys.* **1992**, *96*, 8962–8970.
- Mo, Y. R.; Gao, J. L.; Peyerimhoff, S. D. *J. Chem. Phys.* **2000**, *112*, 5530–5538.
- Mitoraj, M. P.; Michalak, A.; Ziegler, T. *J. Chem. Theory Comput.* **2009**, *5*, 962–975.
- Wu, Q.; Ayers, P. W.; Zhang, Y. K. *J. Chem. Phys.* **2009**, *131*, 164112.
- Donchev, A. G.; Ozrin, V. D.; Subbotin, M. V.; Tarasov, O. V.; Tarasov, V. I. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 7829–7834.
- Piquemal, J. P.; Chevreau, H.; Gresh, N. *J. Chem. Theory Comput.* **2007**, *3*, 824–837.

- (58) Wu, J. C.; Piquemal, J. P.; Chaudret, R.; Reinhardt, P.; Ren, P. Y. *J. Chem. Theory Comput.* **2010**, *6*, 2059–2070.
- (59) Mayer, I. *Phys. Chem. Chem. Phys.* **2006**, *8*, 4630–4646.
- (60) Reinhardt, P.; Piquemal, J. P.; Savin, A. *J. Chem. Theory Comput.* **2008**, *4*, 2020–2029.
- (61) Mo, Y. R.; Bao, P.; Gao, J. L. *Phys. Chem. Chem. Phys.* **2011**, *13*, 6760–6775.
- (62) Mo, Y. R.; Song, L. C.; Lin, Y. C. *J. Phys. Chem. A* **2007**, *111*, 8291–8301.
- (63) Steinmann, S. N.; Corminboeuf, C.; Wu, W.; Mo, Y. R. *J. Phys. Chem. A* **2011**, *115*, 5467–5477.
- (64) Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098–3100.
- (65) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648–5652.
- (66) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. *J. Chem. Phys.* **2010**, *132*, 154104.
- (67) Lee, C. T.; Yang, W. T.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785–789.
- (68) Wu, Q.; Yang, W. T. *J. Chem. Phys.* **2002**, *116*, 515–524.
- (69) Zhao, Y.; Truhlar, D. G. *Theor. Chem. Acc.* **2008**, *120*, 215–241.
- (70) Zhao, Y.; Truhlar, D. G. *Acc. Chem. Res.* **2008**, *41*, 157–167.
- (71) Thanthiriwatte, K. S.; Hohenstein, E. G.; Burns, L. A.; Sherrill, C. D. *J. Chem. Theory Comput.* **2011**, *7*, 88–96.
- (72) Hujó, W.; Grimme, S. *Phys. Chem. Chem. Phys.* **2011**, *13*, 13942–13950.
- (73) Riley, K. E.; Pitonak, M.; Jurecka, P.; Hobza, P. *Chem. Rev. (Washington, DC, U. S.)* **2010**, *110*, 5023–5063.
- (74) Riley, K. E.; Pitonak, M.; Cerny, J.; Hobza, P. *J. Chem. Theory Comput.* **2010**, *6*, 66–80.
- (75) Stone, A. J. *J. Chem. Theory Comput.* **2005**, *1*, 1128–1132.
- (76) Cisneros, G. A.; Tholander, S. N. L.; Parisel, O.; Darden, T. A.; Elking, D.; Perera, L.; Piquemal, J. P. *Int. J. Quantum Chem.* **2008**, *108*, 1905–1912.
- (77) Piquemal, J. P.; Gresh, N.; Giessner-Prettre, C. *J. Phys. Chem. A* **2003**, *107*, 10353–10359.
- (78) Slipchenko, L. V.; Gordon, M. S. *J. Comput. Chem.* **2007**, *28*, 276–291.
- (79) Wang, B.; Truhlar, D. G. *J. Chem. Theory Comput.* **2010**, *6*, 3330–3342.
- (80) Halgren, T. A. *J. Am. Chem. Soc.* **1992**, *114*, 7827–7843.
- (81) Grimme, S. *J. Comput. Chem.* **2006**, *27*, 1787–1799.
- (82) Jurecka, P.; Cerny, J.; Hobza, P.; Salahub, D. R. *J. Comput. Chem.* **2007**, *28*, 555–569.
- (83) Xu, X.; Goddard, W. A. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 2673–2677.
- (84) Boys, S. F.; Bernardi, F. *Mol. Phys.* **1970**, *19*, 553–&.
- (85) Wu, Q.; Yang, W. T. *J. Chem. Phys.* **2003**, *118*, 2498–2509.
- (86) Zhang, Y. K.; Pan, W.; Yang, W. T. *J. Chem. Phys.* **1997**, *107*, 7921–7925.
- (87) Wu, Q.; Van Voorhis, T. *Phys. Rev. A* **2005**, *72*, 024502.
- (88) Cohen, A. J.; Mori-Sanchez, P.; Yang, W. T. *Science* **2008**, *321*, 792–794.
- (89) Zhang, Y. K.; Yang, W. T. *J. Chem. Phys.* **1998**, *109*, 2604–2608.
- (90) Breneman, C. M.; Wiberg, K. B. *J. Comput. Chem.* **1990**, *11*, 361–373.
- (91) Bayly, C. I.; Cieplak, P.; Cornell, W. D.; Kollman, P. A. *J. Phys. Chem.* **1993**, *97*, 10269–10280.
- (92) Ponder, J. W.; Wu, C. J.; Ren, P. Y.; Pande, V. S.; Chodera, J. D.; Schnieders, M. J.; Haque, I.; Mobley, D. L.; Lambrecht, D. S.; DiStasio, R. A.; Head-Gordon, M.; Clark, G. N. I.; Johnson, M. E.; Head-Gordon, T. *J. Phys. Chem. B* **2010**, *114*, 2549–2564.
- (93) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; J. R. Cheeseman, 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.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03*, revision B.05; Gaussian, Inc.: Pittsburgh PA, 2003.
- (94) Ponder, J. W. *TINKER, Software Tools for Molecular Design*, version 5.0; Washington University: St. Louis, MO, 2009.
- (95) Bondi, A. *J. Phys. Chem.* **1964**, *68*, 441–450.
- (96) Kendall, R. A.; Apra, E.; Bernholdt, D. E.; Bylaska, E. J.; Dupuis, M.; Fann, G. I.; Harrison, R. J.; Ju, J. L.; Nichols, J. A.; Nieplocha, J.; Straatsma, T. P.; Windus, T. L.; Wong, A. T. *Comput. Phys. Commun.* **2000**, *128*, 260–283.
- (97) Johnson, E. R.; Wolkow, R. A.; DiLabio, G. A. *Chem. Phys. Lett.* **2004**, *394*, 334–338.
- (98) Johnson, E. R.; Becke, A. D.; Sherrill, C. D.; DiLabio, G. A. *J. Chem. Phys.* **2009**, *131*, 34111.
- (99) Wheeler, S. E.; Houk, K. N. *J. Chem. Theory Comput.* **2010**, *6*, 395–404.
- (100) Becke, A. D. *J. Chem. Phys.* **1988**, *88*, 2547–2553.
- (101) Gordon, M. S.; Freitag, M. A.; Bandyopadhyay, P.; Jensen, J. H.; Kairys, V.; Stevens, W. J. *J. Phys. Chem. A* **2001**, *105*, 293–307.
- (102) Louwen, J. N.; Vogt, E. T. C. *J. Mol. Catal. A: Chem.* **1998**, *134*, 63–77.
- (103) Elking, D. M.; Cisneros, G. A.; Piquemal, J. P.; Darden, T. A.; Pedersen, L. G. *J. Chem. Theory Comput.* **2010**, *6*, 190–202.

Solvent Effects on the Radiative and Nonradiative Decay of a Model of the Rhodopsin Chromophore

Aurora Muñoz-Losa,* M. Elena Martín, Ignacio Fdez. Galván, M. Luz Sánchez, and Manuel A. Aguilar

Química Física, Universidad de Extremadura, Avda. de Elvas s/n 06071 Badajoz, Spain

S Supporting Information

ABSTRACT: The radiative and nonradiative decay of a model with five double bonds of the 11-*cis*-retinal protonated Schiff base was studied both in vacuum and in methanol solution using an extended version of the averaged solvent electrostatic potential from molecular dynamics data (ASEP/MD) method that allows the location of crossing points between free energy surfaces both in equilibrium and in frozen solvent conditions. The multireference quantum method CASSCF was used for the description of the states of interest, while the solvent structure was obtained from molecular dynamics simulations. Electron dynamic correlation corrections to the energy were included at CASPT2 level. Unlike in gas phase, where only two states seem to be implicated, in methanol solution, three states are necessary to describe the photoisomerization process. At the Franck–Condon point the S_1 and S_2 states are almost degenerate; consequently, the S_1 surface has a region with an ionic character (1B_u -like) and another one with a covalent character (2A_g -like). Emission from the ionic minima is responsible for the low-frequency part of the fluorescence band, while emission from the covalent minima originates the high-frequency part. The ionic minimum is separated from the conical intersection yielding the all-*trans* isomer by an energy barrier that was estimated in 0.7 kcal/mol. The geometry of the optimized conical intersection was found at a torsion angle of the central double bond close to 90° both in vacuum and in methanol solution. This large torsion in addition to the accompanying charge displacements forces a strong solvent reorganization during the de-excitation process which slows down the photoisomerization kinetics in methanol with respect to the gas phase. Solvent fluctuations modulate the minima depth and the barrier height and could explain the multiexponential relaxation time observed in the experiments.

1. INTRODUCTION

Although the theoretical study of solvent effects on the evolution of molecules in excited states has received an increasing attention in the last years, nonradiative (nonadiabatic) processes have been comparatively less studied than radiative processes (fluorescence or phosphorescence). The reasons are obvious, from the difficulties inherent to the presence of a solvent, as are the existence of a manifold of configurations thermally accessible, to the great number of solvent molecules involved or the interplay between solute and solvent dynamics, one must add the complications associated to the study of nonadiabatic processes,^{1–5} processes that imply more than one potential energy surface. It is usual to classify nonadiabatic processes as internal conversion (IC) or intersystem crossing (ISC) depending on the spin symmetry of the states involved. Associated with this nonadiabatic process we find a potential energy surface crossing named conical intersection (CI) or singlet–triplet crossing (STC). Nowadays, there are several techniques and algorithms available that permit the determination of those geometries for which CI and STC appear and also for the minimal energy conical intersection geometries (MECI), which are considered as the most probable radiationless decay sites.

There have been several proposals in the literature to include solvent effects in the study of IC processes. Burghardt et al.,^{6,7} for instance, use dielectric continuum methods. These authors introduce an explicit coordinate for the solvent, which permits them to study the solvation dynamics during the internal

conversion process. Polarizable continuum models (PCM) have been used by Barone et al. in the study of uracil derivatives.⁸ Methods that allow a more detailed description of the solvent have also been proposed; so for instance, Yamazaki and Kato⁹ use the reference interaction site model self-consistent field (RISM-SCF)¹⁰ method for describing the solvent dynamics during energy surface crossing in ethylene and CH_2NH_2^+ in polar solvents. This group has also studied conical intersections in a small protonated Schiff base model of retinal in methanol solution, stressing the importance of the electron correlation in the structural and energetic properties of the MECI.¹¹ Other groups^{12–14} have used quantum mechanics/molecular mechanics (QM/MM) methods to locate CI, generally in frozen solvent conditions, although in some cases the solvent dynamics has been also considered.

In a previous study¹⁵ we developed an extended version of the averaged solvent electrostatic potential from molecular dynamics (ASEP/MD) method that permitted the study of the solvent effects on radiationless decay processes both in equilibrium and nonequilibrium (frozen) solvent conditions. As an application of the method, we studied the photophysics of acrolein, solvent reorganization was not a necessary condition for the photoisomerization reaction, and the molecule exhibited a similar behavior in solution and in gas phase. In the present study we are interested in the location and characterization of the principal critical points

Received: April 28, 2011

Published: November 04, 2011

(minima, CI, etc) on the first two low-lying excited states of a model of the rhodopsin chromophore in methanol solution.

Rhodopsin is a protein that is highly specialized in the detection of photons. Its chromophore, retinal, is an A1-vitamin derivative and is formed by a β -ionone ring and a polyene chain bonded covalently through a Schiff bond to the Lys296 residue of the protein. As a consequence of the absorption of one photon, the protonated Schiff base of the 11-*cis*-retinal (PSB11) undergoes a rotation, and it transforms into its all-*trans* isomer in a very fast process that, inside the protein, happens in less than 200 fs.^{16,17} The speed of the process and the lack of radiation emission is explained by the existence of a conical intersection between the potential energy surfaces of the first excited state and the ground state. These states are clearly differentiated by its charge distribution, that is, the ground state has a predominantly covalent character (dot–dot) with a localized electronic distribution, whereas the excited crossing state has a predominantly ionic character (hole–pair), which means that the charges are delocalized over the molecule. Several studies of this nonadiabatic process have been carried out. Martínez et al.,^{12b} for instance, used the floating occupation molecular orbital approach (FOMO) with semiempirical methods to locate the MECI in vacuum, although the energies are not in agreement with those obtained with *ab initio* calculations. In a previous work^{12a} they used a QM/MM method to solvate an analogue of PSB11 with 57 MM water molecules; they found that the stabilization of the MECI with respect to in vacuum conditions is about 7 kcal/mol. Olivucci et al. have largely studied different aspects of this IC, they have analyzed different models of the PSB11 in vacuum,^{13,18} the effects of the counterions,¹⁹ and more recently, the effect of the opsin.^{13,20} Burghardt et al.⁶ use a dielectric continuum model to describe the electrostatic effects of the environment in protonated Schiff bases, like PSB11, concluding that the CI is lost in frozen solvent conditions. Send and Sundholm²¹ carried out a time-dependent density functional theory (TD-DFT) study in vacuum and inside the protein; they suggested that the electron excitation produces the torsion of the β -iononic ring, and then the torsion is propagated along the carbon chain to the C₁₁=C₁₂ center to allow the photoisomerization. Warshel and Chu²² analyze the nature of the surface crossing process in bacteriorhodopsin using a hybrid QM/MM method. It is demonstrated that the motion starts with bond vibrations and evolves to a torsional motion and that surface crossing occurs only in the 90° region. More recently, Kato et al. have published a theoretical study of the *cis*–*trans* photoisomerization of a small PSB11 model with only three double bonds in methanol solution.¹¹ These authors employed analytical gradients for the location of the MECI at multistate-CASPT2 level²³ and the RISM-SCF method¹⁰ to include the solvent effects. In this study the importance of the dynamical correlation in the structural and energetic properties of two possible MECI has been highlighted.

In a more experimental vein, several authors^{24–27} analyzed the photoisomerization process inside the protein, and evidence for the involvement of a conical intersection was obtained. In a recent study, Zgrablić et al.²⁸ have obtained femtosecond fluorescence spectra of the all-*trans* retinal Schiff base in several polar and apolar solvents. The analysis of the time-resolved fluorescence spectrum in methanol reveals a complex spectral behavior that could be originated by emissions from different structures and states. Other experimental data to account for is that, in methanol, the photoisomerization is 2 orders of magnitude slower than in the protein, while calculations in gas phase report a process as fast as in protein.²⁹ Elucidating if this slow reaction in

methanol is due to changes in the free energy surfaces of the states involved in the photoisomerization or if, on the contrary, it is related to solvent dynamics is one of the aims of this paper. It is worth noting that during the *cis*–*trans* photoisomerization of retinal, there are parts of the solute molecule that can suffer large displacements, consequently, it can be expected that, in opposition to what was found in acrolein, the solvent reorganization would become an important step in the reaction mechanism and lead to large differences between the gas phase and in solution behavior. A second aim is the location of the possible minima that could explain the fluorescence spectrum of the chromophore in methanol. In short, we expect to clarify the role that the solvent plays in the radiative and nonradiative decay processes.

The rest of the paper is organized as follows: In Section 2 the main features of the method will be explained, and the computational aspects will be detailed. In Section 3 the in vacuum and in methanol solution results will be discussed, and finally, the main conclusions will be reported in Section 4.

2. METHODS AND DETAILS

ASEP/MD is a QM/MM effective Hamiltonian method that makes use of the mean field approximation.³⁰ The method combines QM and MD techniques with the particularity that full QM and MD calculations are alternated and not simultaneous. During the MD simulations, the intramolecular geometries and charge distributions of all molecules are considered as fixed, and the interaction is calculated with MM. From the resulting data, the average electrostatic potential generated by the solvent on the solute can be obtained. This potential is introduced as a perturbation into the solute's quantum mechanical Hamiltonian, and by solving the associated Schrödinger equation, a new charge distribution for the solute is obtained, which is used in the next MD simulation. The iterative process is repeated until the electron charge distribution of the solute and the solvent structure around it become mutually equilibrated.

To locate a optimized conical intersection between two electronic states, we combine the ASEP/MD method with an algorithm due to Bearpark et al.³¹ The algorithm simultaneously minimizes the in solution energy difference between the two intersecting states and the energy of the crossing seam between the two potential energy surfaces. Details of the method can be found elsewhere.¹⁵

We consider two possible regimes for the solvent depending on whether it is in an equilibrium or nonequilibrium (frozen) situation with the solute. In the former case, the solvent is equilibrated at every point with the solute charge distribution of the adiabatic state on which the initial dynamics takes place, i.e., the S₁ state with ionic character when an S₁/S₀ CI is searched for. Although in the neighborhood of the CI, the S₀ and S₁ states are swapping, the solvent is always in equilibrium with the state of larger ionic character, which is identified by its charge distribution. In the second regime, a frozen solvent structure, which is in equilibrium with the charge distribution of a previous solute structure, is used. In this situation the solvent does not respond to changes in the solute charge distribution. Both situations are extreme cases, the real behavior of the system is expected to be somewhere between them. It is worth noting that the minimum energy conical intersection located in these conditions is an upper limit to the real MECI; in consequence, we do not obtain the absolute MECI but the lowest energy CI subject to the condition that the solvent is either in equilibrium with the S₁ state

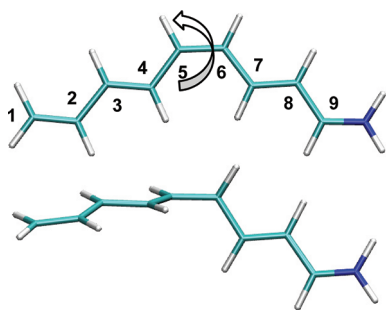


Figure 1. Planar M1 and twisted EQ-MECI M1 pictures and numbering of the carbon skeleton.

or frozen. Therefore from now on we call this optimized conical intersection in solution as EQ-MECI in equilibrium and FS-MECI with frozen solvent.

Once the geometries of interest (ground- and excited-state minima, EQ-MECI, etc.) have been located the free energy differences, ΔG , between them are calculated as sums of two terms

$$\Delta G = \Delta E_{QM} + \Delta G_{int} \quad (1)$$

where ΔE_{QM} is the difference in the internal quantum energy of the solute between the two geometries and ΔG_{int} is the difference in the solute–solvent interaction energy, which is calculated classically using the free energy perturbation method.³² In turn, the ΔG_{int} term can be split into two terms $\Delta G_{int} = \Delta E_{int} + \Delta G_{solv}$. The term ΔE_{int} accounts for the difference in the solute–solvent interaction energy between the final and initial state. The last term, ΔG_{solv} , provides the solvent distortion energy, i.e., the energy spent in changing the solvent structure from the initial to the final state.

2.1. Computational Details. The current study tackles the comparison of the excited potential energy surfaces of a model of the 11-*cis*-retinal protonated Schiff base molecule formed by five double bonds in vacuum and in methanol solution. In previous papers³³ it was shown that this model, called M1 (see Figure 1), is adequate in studying the photophysical behavior of the real molecule since it reproduces some of the main features of the experimental absorption spectra:³⁴ two well separated excited states in vacuum that become almost degenerate in methanol solution. The ground and excited states were described using state-average complete active space self-consistent field (SA-CASSCF) of the first three roots with equal weights. All electrons of the π skeleton were included in the active space, which was spanned by all the configurations arising from 10 valence π electrons in 10 orbitals (10e, 10o). The split-valence 6-31G(d) basis set was employed. It is well-known that to obtain an accurate description of the energetic properties of photoexcited systems, the dynamic electronic correlation must be included. In our case, we used the second-order perturbation method CASPT2. Since analytical gradients are not available for this method in the MOLCAS program, carrying out CASPT2 geometry optimizations was impractical, and therefore we only recalculated the electronic energies with CASPT2 at the geometries located with SA-CASSCF. In the case of the MECI, also multistate CASPT2²³ calculations were performed. All the calculations were performed with a development version of the ASEP/MD program using the data provided by Gaussian98³⁵ and Moldy.³⁶ The dynamical correlation corrections were calculated

with MOLCAS 6.4.³⁷ Calculations were performed with no ionization potential electron affinity (IPEA) shift³⁸ to be consistent with previous calculations done with older MOLCAS versions. An additional imaginary shift of $0.1i E_h$ was included in order to minimize the appearance of intruder states. All the minima, both in vacuum and in methanol solution, were confirmed by analytical Hessian calculations at CASSCF with a harmonic approximation and, in the case of in methanol minima, supposing that the solvent remains frozen during the solute vibration.

To locate the CI points, we used a quasi-Newton method where the approximate Hessian was updated by using the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm.

A total of 331 molecules were simulated with fixed intramolecular geometry by combining Lennard-Jones interatomic interactions with electrostatic interactions. The solvent was represented by 330 molecules of methanol using AMBER nonbonded parameters³⁹ in a cubic box of 28.2 Å side (test calculations performed with 1000 methanol molecules did not show significant changes). Also AMBER nonbonded parameters³⁹ were used for the solute. No counterion was included. Previous studies of Rajamani and Gao⁴⁰ and Röhrig et al.⁴¹ using chloride as counterion find that, because of the large dielectric screening effect of methanol, the effect of the counterion on the structure and spectrum of the solute is minimal. This has been corroborated by experiments showing that the position of the chromophore absorption band in polar solvents is not affected by the nature of the counterion.⁴² Periodic boundary conditions were applied, and spherical cutoffs were used to truncate the molecular interactions at 9.0 Å. A time step of 0.5 fs was used. The electrostatic interaction was calculated with the Ewald method. The temperature was fixed at 298 K by using a Nosé–Hoover thermostat. Each MD calculation simulation was run for 75 ps (25 ps equilibration, 50 ps production). In solution, the results are affected by statistical uncertainty due to the finite size of the MD sampling, and we take average values of the last five ASEP/MD cycles. Given that 10–15 total cycles are performed, the effective times are around 400–750 ps for equilibration and 250 ps for production.

3. RESULTS AND DISCUSSION

In previous papers³² it was shown that the solvent modifies the relative stability of the low-lying excited states of the 11-*cis*-retinal protonated Schiff base and of several molecules used as models of the rhodopsin chromophore, M1 among them. In gas phase the absorption spectrum displays two bands separated by almost 1 eV, corresponding to transitions from the ground state (covalent or 2A_g -like in character) to the first excited state, which has ionic character (1B_u -like) and to the second excited state, with a covalent character. The transition to the covalent excited state is generally associated to a smaller oscillator strength and can be difficult to detect experimentally in some cases. In methanol solution a single band is observed in the absorption spectrum.

In the next two subsections we will describe the main characteristics of the two low-lying excited states of M1 both in gas phase (Section 3.1) and in methanol solution (Section 3.2). We will show that in both phases, it has been possible to locate several local minima on S_1 . We will also discuss the solvent effects on the relative stability of the EQ-MECI. A priori, any double bond of the 11-*cis*-retinal molecule could undergo photoisomerization; in fact, in solution, conical intersections have been identified^{12b} that lead to the all-*trans* retinal and to several di-*cis*

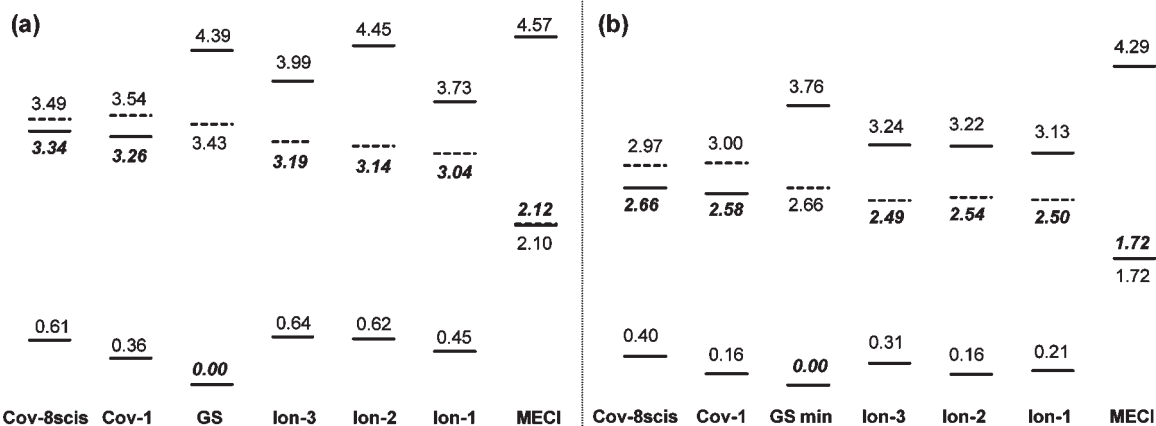


Figure 2. Relative energies (in eV) with respect to the in equilibrium ground state at CASSCF level (a) and at CASPT2//CASSCF level (b) in vacuum for the GS, Ion-1, Ion-2, Ion-3, Cov-1, and Cov-8scis minima, and the MECI. Covalent states are in solid lines, and ionic states are in dashed lines. In bold and italic, the optimized state energies.

isomers. The process is more selective inside the rhodopsin pocket where the only bond that undergoes the rotation is the central double bond $C_{11}=C_{12}$. The final product in this case is the all-*trans* retinal. It has been suggested^{19,43} that the presence of a counterion from the protein can favor the isomerization of this particular bond. In our study we mainly focus on the photoisomerization of the central double bond, although other possibilities are commented.

3.1. In Gas Phase. Figure 2 displays the gas phase transition energies computed at CASSCF and CASPT2 levels. For all the structures considered (FC, minima, MECI), the ionic excited state is more stable than the excited covalent state, except in the covalent-state minima. So, for instance, in the FC point the first two excited states are very well-defined and separated, about 1 eV independently of the calculation level, CASSCF or CASPT2, used. The same is valid in the MECI region. As has been suggested in previous studies^{18b} and our calculations confirm, the covalent excited state is not directly involved in the conical intersection associated to the internal conversion process in vacuum. Only two states, ground and ionic, are involved in the photoisomerization reaction. The main structural parameters of the minima located in gas phase are gathered in Table 1. At the beginning of the absorption process, the molecule has the planar structure of the ground state with a clear alternation of single and double bond lengths. Thus, the bond length alternation (BLA), that is, the difference between the length of formal simple and double bonds, is 0.39 Å. The main characteristic of the potential energy surface of the ionic excited state is the presence of several local minima and of a CI. The minima were located at 2.50, 2.54, and 2.49 eV with respect to the ground-state minimum at CASPT2//CASSCF level and only around 0.15 eV under the FC point. These minima, called from now on Ion-1, Ion-2, and Ion-3, respectively, are very close in energy, and all of them display a planar structure. The Ion-1 minimum has a single- and double-bond distribution opposite to that displayed by the ground state: the BLA value is -0.14 Å (the minus sign indicates the interchange in the nature of single and double bonds). In the Ion-2 minimum all the central bonds have the same length, and consequently, the BLA value is 0.00 Å. Finally, the Ion-3 minimum displays a large bond alternation which translates in a BLA value of -0.24 Å. It is worth noting that the three ionic

Table 1. Bond Lengths (in Å) for Ground and the Different Minima of the Excited States and the MECI Points in Vacuo

	GS	Ion-1	Ion-2	Ion-3	Cov-1	MECI
C_1C_2	1.35	1.37	1.36	1.37	1.42	1.36
C_2C_3	1.46	1.42	1.41	1.40	1.38	1.42
C_3C_4	1.35	1.42	1.42	1.43	1.42	1.41
C_4C_5	1.45	1.38	1.41	1.36	1.40	1.38
C_5C_6	1.36	1.46	1.42	1.48	1.42	1.47
C_6C_7	1.45	1.37	1.41	1.36	1.40	1.37
C_7C_8	1.35	1.44	1.42	1.46	1.41	1.42
C_8C_9	1.44	1.38	1.39	1.38	1.41	1.39
C_9N	1.28	1.33	1.33	1.34	1.32	1.32

minima differ from each other in less than 1.0 kcal/mol. The three structures are local minima confirmed through frequency calculations, and we found several low-frequency modes corresponding to movements in the molecular plane, indicating that the surface between the minima is relatively flat. The presence of these many minima could be an artifact of using CASSCF method in the optimization. In fact, a previous study by Valsson and Filippi,⁴⁴ several minima were found at CASSCF level, whereas a single one was found with other methods that include dynamical correlation. The minimal energy CI is placed at a dihedral angle $C_4C_5C_6C_7$ twist of 90° with an energy that is 1.72 eV above the ground-state minimum, i.e., the MECI is almost 0.80 eV more stable than the ionic minima and 0.94 eV lower than the FC point. This MECI was confirmed by the MS-CASPT2 method with an energy above the ground-state minimum of 1.73 eV. We find that the ionic minima are separated from the MECI by a very small barrier with a height lower than 1 kcal/mol. The barrier height was estimated through a linear interpolation in internal coordinates between the geometries of the ionic S_1 minimum and MECI, and the obtained value can only be considered as an upper limit, subject to all the approximations of the calculation. In these conditions one expects that most of M1 molecules decay through the CI, consequently a low value of the fluorescence quantum yield is expected.

The MECI geometry is characterized by the great torsion of the $C_4C_5C_6C_7$ dihedral angle and by the increase of the central double-bond length C_5C_6 up to 1.47 Å, 0.11 Å larger than at the ground-state minimum geometry. The rest of the molecule remains planar even when we start the optimization from distorted geometries. The loss of the π character of the C_5C_6 central bond eases its torsion and permits the rotation of the molecule. During the evolution of the system from the FC point to the MECI, the differences between double- and single-bond distances decrease, and the BLA in the MECI becomes -0.10 Å, which has been described in previous works.⁴⁵ Other mechanisms for the isomerization are possible, and in particular, a bicycle-pedal mechanism has been proposed to occur inside the protein cavity, where spatial constraints are severe,⁴⁶ and it has also been observed in *ab initio* trajectories in gas phase.⁴⁷ We tried to locate other MECI points corresponding to different isomerization mechanisms, including the bicycle-pedal, but the search was not successful in this system.

We also found a planar minimum, named Cov-1, on the covalent excited state. All the bonds display similar bond lengths, consequently the BLA value is very small, only -0.08 Å. This minimum (see Figure 2) is only 0.09 eV above the ionic state minimum. Other minima with energies similar to Cov-1 were found on the covalent excited state by rotation around single bonds, one of these minima denoted as Cov-8scis is also displayed in Figure 2.

From the above results we can conclude that most molecules decay without radiation emission through the CI that connects the ionic and the ground state. The possible fluorescence signal would originate mainly at the ionic state minima, while the contribution of the covalent excited state would be minor, owing to the value of the oscillator strength, which is between 1.0 and 1.6 for the ionic minima and only 0.04 for the covalent one. The presence of several minima points to a complex spectral behavior characterized by several relaxation times. The CASPT2 emission energies are 2.29, 2.38, and 2.17 eV for Ion-1, Ion-2 and Ion-3, respectively, that are somewhat lower than the energy obtained at TD-CAM-B3LYP level, 2.46 eV.⁴⁸

A limitation of our study is that the excited states geometries have been optimized at CASSCF level. Recent papers of Valsson et al.⁴⁴ and Yamazaki and Kato⁹ have highlighted the importance that the inclusion of the dynamic correlation could have on the optimized geometries of excited states. For M1 we found that the CI is kept when the energies are recalculated at CASPT2 level or MS-CASPT2, the energy difference between the two crossing states being lower than 0.1 kcal/mol. This result agrees with that obtained by Valsson et al.,⁴⁴ who concluded that the structures calculated at CASSCF and CASPT2 level are similar in the vicinity of the conical intersection. More important differences are expected at the minima geometries, and in this case, it has been shown that CASSCF tends to overestimate the BLA.^{15,44} Translating this trend to M1, it is probable that the different ionic state minima located at CASSCF level would have much closer geometries or even collapse in a single minimum at CASPT2 level; in any case, the similarity in CASPT2 energies for Ion-1, Ion-2 and Ion-3 indicates that the CASPT2 surface is still relatively flat, which can still lead to a complex relaxation behavior.

3.2. In Methanol Solution. With the absorption of one photon, rhodopsin undergoes the isomerization of the 11-*cis*-retinal to all-*trans* form in a very fast process that takes less than 200 fs.^{16,17} In methanol the isomerization process is 2 orders of magnitude slower, taking 10 ps for the transformation to the

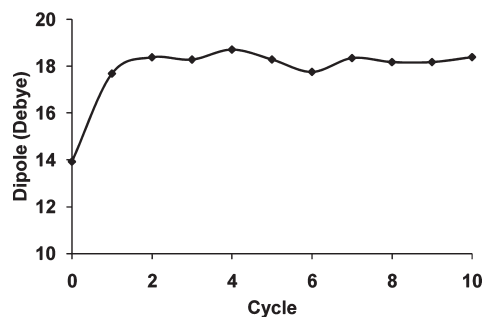


Figure 3. Evolution of the dipole moment in debye during the ground-state optimization vs the number of ASEP/MD cycles employed.

all-*trans* isomer.¹⁶ In this case a transient fluorescent state is formed with a 3 ps fluorescence lifetime, whereas inside the protein this state fluoresces only for 50–60 fs.^{17,49} In a recent experimental study, Zgrablić et al.²⁸ disentangle the different spectro-temporal components that make up the fluorescence spectra. In methanol, three spectral components are isolated that Zgrablić et al. interpret as associated to: (1) the vibrationally relaxed S_1 fluorescence, (2) a vibrationally hot S_1 fluorescence, and (3) a higher-lying emission which was assigned to S_2 fluorescence. Trying to relate these experimental data with the characteristics of the free energy surface of the first two excited states and with the solvent response is the goal of this section.

All the states have been optimized in methanol with the ASEP/MD method. The results were obtained as the average of the last five cycles, where the system properties are converged. Figure 3 displays the evolution of the dipole moment during the ASEP/MD procedure. Fluctuations are due to the limited size of the dynamics, as have been analyzed in previous works.⁵⁰

Figure 4 displays the relative stability of the different states of M1 in presence of methanol at CASSCF and CASPT2 level, respectively. All the minima have been obtained in equilibrium solvent conditions and confirmed as such with frequency calculations, while in the Franck–Condon points we have assumed frozen solvent conditions during the transition. Because of its importance in the solvation process, it is interesting to compare the electronic densities of covalent and ionic states, see Figure 5. As a general trend, covalent and ionic excited states are destabilized with respect to the ground state, although not in the same extent. In the ionic state the molecular charge is delocalized along the whole molecule, consequently it is worse solvated than covalent states where the charge is localized in the iminium end. The ionic state destabilization is especially important at the Franck–Condon point at CASPT2 level. We tried to validate this result with MS-CASPT2 calculations, but the resulting off-diagonal terms of the Hamiltonian matrix were too large for the results to be trusted;⁵¹ we enlarged the active space up to 16 orbitals, and this problem was not solved. More details can be found in the Supporting Information. In this situation, where the CASSCF functions are not degenerate and the MS-CASPT2 calculation yields unphysical results, the usual advice is to rely on the single-state CASPT2 energies, and therefore, with appropriate prudence, we take the CASPT2 results at the Franck–Condon point as good. It is worth noting that, as displayed in Figure 4, in methanol, ionic and covalent excited states cross in the neighborhood of the FC point, consequently part of the S_1 surface has an ionic character and part covalent. In order to check the crossing between the two

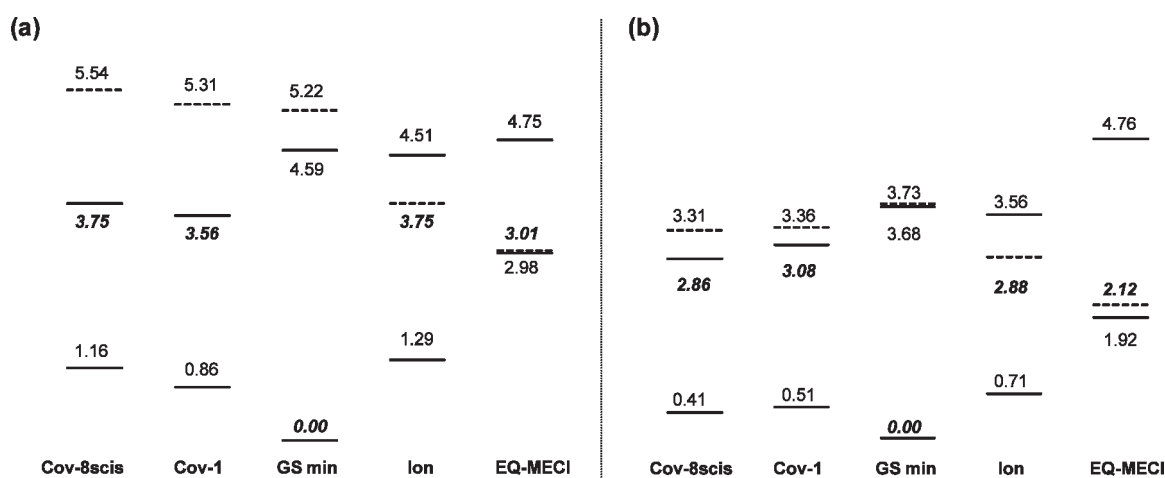


Figure 4. Relative free energies (in eV) with respect to the in equilibrium ground state at CASSCF level (a) and at CASPT2//CASSCF level (b) in methanol for the GS, Ion, Cov-1 and Cov-8scis minima, and the EQ-MECI. Covalent states are in solid lines, and ionic states are in dashed lines. In bold and italic, the optimized state energies.

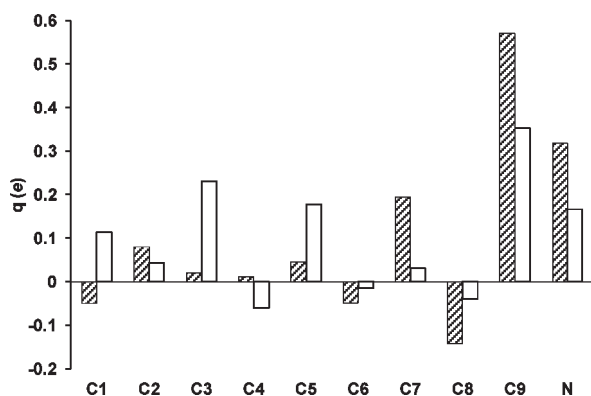


Figure 5. Charges (in e , fitted to the CASSCF electrostatic potential) in methanol for the ionic excited state (white columns) and ground state (hatched columns). Charges for the covalent excited state are similar to the ground state.

excited states, 50 configurations were extracted from the molecular dynamics (MD) with the solvent in equilibrium with the ground state, next the electronic transitions were calculated. It was found that the covalent state is below the ionic state in approximately one-half of the configurations. It is observed, too, that the electronic transition to the ionic state remains quasi-invariable, with a fluctuation of 0.02 eV, measured as the standard deviation, whereas the covalent state undergoes larger fluctuations of 0.11 eV.

Regarding the ionic state, a point to remark is that in methanol solution, and independently of the geometry taken as a starting structure (Ion-1, Ion-2, or Ion-3), it was possible to locate only one ionic minimum on S_1 with a relatively flat surface, as in gas phase. This minimum is 0.85 eV under the FC point. The experimental fluorescence energy due to the emission from the ionic excited state is 1.72 eV,²⁸ whereas our calculated value is 2.17 eV. This last value is almost the same as found in gas phase. The oscillator strength is close to 1. The overestimation of the calculated transition energy is due to: (1) the lower number of conjugated double bonds of M1 (five) with respect to retinal (six) and (2) the nature of the groups bonded to the nitrogen

Table 2. Free Energy Differences and Their Components (in eV) between Pairs of Critical Points at CASSCF Level (in parentheses at CASPT2//CASSCF) in Methanol Solution

	ΔE_{QM}	ΔE_{int}	ΔG_{solv}	ΔG_{int}	ΔG
GS (S_0) \rightarrow EQ-MECI	2.48	1.75	-1.21	0.54	3.01 (2.12)
GS (S_1) \rightarrow EQ-MECI	-1.83	0.85	-1.21	-0.37	-2.21 (-1.60)
GS (S_0) \rightarrow Ion	3.08	1.21	-0.55	0.67	3.75 (2.88)
GS (S_1) \rightarrow Ion	-1.24	0.31	-0.55	-0.23	-1.47 (-0.87)
GS (S_0) \rightarrow Cov-1	3.56	-0.11	0.11	0.00	3.56 (3.08)
GS (S_2) \rightarrow Cov-1	-0.91	-0.23	0.11	-0.12	-1.03 (-0.60)

atom in the iminium group. In M1 the N atom is bonded to two hydrogen atoms, while the experiments have been performed with molecules where the N is bonded to $-\text{CH}_3$ or more bulky groups. Consequently, M1 provides larger solvation energy values.¹⁵

We also found a covalent minimum on S_1 placed 0.60 eV under the FC point and only 0.2 eV over the ionic minimum. The calculated fluorescence emission from this state is 2.57 eV, and the estimated experimental value is 2.17 eV. The deviation from the experimental value, 0.40 eV, is similar to that obtained for the emission from the ionic state and is due to the same reasons. The oscillator strength of the covalent minimum is only 0.08 because at this geometry, S_1 and S_0 have the same character. From a photochemical point of view, the most remarkable characteristic of the ionic excited-state surface is the presence of the S_0/S_1 conical intersection.⁵ The EQ-MECI point is placed 0.75 eV under the ionic-state minimum. This value is similar to that found in gas phase. Again, the presence of the EQ-MECI was validated by the MS-CASPT2 method. As the molecule twists, the covalent excited state goes away from the ionic excited state, and at the EQ-MECI, geometry is 2.64 eV above the ionic state. The S_0/S_1 EQ-MECI therefore does not involve the covalent excited state.

Table 2 displays the contributions to the relative free energy differences. The most important contribution is ΔE_{QM} , i.e., the change in the internal energy as a consequence of the solute distortion, that is, changes in the solute geometry and electronic distribution during the photoisomerization. The solvent contribution

Table 3. Bond Lengths (in Å) for the Minima of the Ground and Excited States and the EQ-MECI in Methanol Solution

	GS	Ion	Cov-1	EQ-MECI
C ₁ C ₂	1.34	1.38	1.43	1.36
C ₂ C ₃	1.47	1.40	1.38	1.40
C ₃ C ₄	1.36	1.43	1.44	1.41
C ₄ C ₅	1.45	1.36	1.39	1.36
C ₅ C ₆	1.36	1.49	1.44	1.46
C ₆ C ₇	1.46	1.35	1.38	1.41
C ₇ C ₈	1.36	1.47	1.44	1.39
C ₈ C ₉	1.43	1.38	1.41	1.44
C ₉ N	1.28	1.33	1.30	1.30

ΔG_{int} is important but only when we compare two states with different charge distributions, the two excited states for instance, but not when the two states have similar charge distribution, ground and covalent excited states. Comparison of ΔE_{QM} with the values displayed in Figure 2 provides the intramolecular contribution to the solvent shift, i.e., the solvent shift due to the change in the solute geometry during the solvation. This contribution is very small in the ionic minimum, 0.11 eV (= 3.19–3.08) and somewhat larger in the covalent minimum, 0.30 eV, and EQ-MECI, 0.35 eV.

Geometrical parameters of the ground- and excited-state minima are gathered in Table 3. In general, the solvent has only a very small effect on the geometry of the different minima. So, for instance, the ground-state geometry is very similar in gas phase, and in methanol, in fact the BLA is the same. The ionic minimum on S_1 is very similar to the Ion-3 gas phase structure with a BLA of -0.28 Å. Similarly to gas phase results, single and double bonds are reverted relative to the ground-state geometry. The S_1 covalent excited-state minimum displays a geometry very similar to the gas phase minimum with only a slight increase of the alternation between single and double bonds providing a BLA value of -0.19 Å. As in gas phase, the minima display planar structures. The solvent effects on the S_0/S_1 EQ-MECI geometry are somewhat larger. As in gas phase, the EQ-MECI structure is characterized by a twisted structure, where the central dihedral angle takes a value of 86° . If we compare the in solution bond lengths with the in vacuum values, one can observe some differences, especially in the iminium end. The rest of the carbon skeleton is hardly affected by the solvent. The BLA in the EQ-MECI geometry is practically zero.

The existence of a stable minimum on S_1 from which the radiative decay is possible points to the existence of a free energy barrier between this minimum and the EQ-MECI. The barrier height was estimated again through a lineal interpolation in internal coordinates between the geometry of the ionic S_1 minimum and the EQ-MECI geometry. Free energy differences were computed every 5° of torsion of the central dihedral. A very small barrier of 0.7 kcal/mol was found with a backbone torsion of 10° . Because of the low height of the barrier, it must be expected that most molecules de-excite through the CI, however, this decay path seems to be less effective in methanol than in gas phase, according to experimental results.¹⁶ In fact, as previously mentioned, theoretical calculations²¹ in gas-phase indicate that the de-excitation through the CI is very fast, around 0.1 ps. On the contrary, in methanol it takes about 10 ps.¹⁶ Given that the barrier height and the relative stability of the ionic minimum and the EQ-MECI are completely similar in gas phase and in methanol solution, the differences in the nonradiative

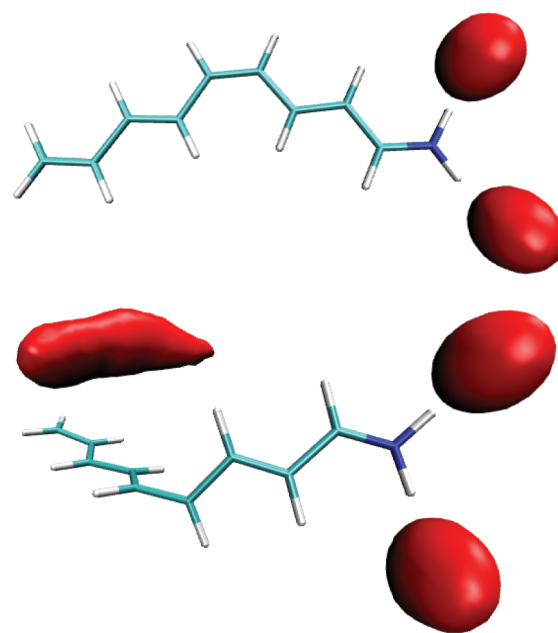


Figure 6. Occupancy maps of methanol oxygen atoms (considered as van der Waals spheres, as calculated by VMD)⁵² around M1 for (a) the optimized ground-state structure and (b) the optimized EQ-MECI structure. Isosurfaces at a value of 0.77.

de-excitation times are probably associated to the solvent dynamics. As we have already indicated, during the *cis*–*trans* photoisomerization of retinal, there are parts of the solute molecule that can suffer large displacements, consequently, it can be expected that the solvent reorganization would become an important step in the nonradiative decay path. In order to clarify the role played by the solvent dynamics, we studied the evolution of the system supposing frozen solvent conditions, i.e., when the solvent is not permitted to relax after the absorption process. Under these conditions our search with the Bearpark algorithm did not locate a low-lying S_1/S_0 CI for the studied process. If the solvent is considered as fixed and in equilibrium with the charge distribution of the S_0 state, the rotation around the C₅C₆ bond during the photoisomerization process, assuming the same simple mechanism found with equilibrated solvent, implies the overlap between the solvent molecules and part of the solute. Because of this steric hindrance, a certain degree of solvent reorganization during the IC process is compulsory or the isomerization mechanism must change. This solvent reorganization is necessary also to stabilize the EQ-MECI. Figure 6 displays the occupancy maps of methanol oxygen atoms around the FC and EQ-MECI structures. At the FC point, the solvent is mainly concentrated around the iminium end that is where most of the charge is localized. In the EQ-MECI, a similar concentration of solvent molecules around the iminium is found, but there is also a high concentration around the carbon skeleton because in this case part of the charge is spread out over the whole molecule. These solvent molecules help to stabilize the charge and hence the EQ-MECI.

The question arises as to whether this difficulty in locating the CI in frozen solvent conditions is due to electrostatic or steric effects. In order to check this, we replaced the solvent representation in ASEP/MD (a set of point charges plus Lennard-Jones

parameters) by a multipole monocentric expansion, centered on the solute molecule, of the electrostatic potential, up to the hexadecapolar field. In any case, the external perturbation represents the average solvent obtained from the MD simulations of the ground state. This alternate solvent representation includes only the electrostatic contribution and discards steric effects. As a validation, the vertical absorption energy obtained with the multipole expansion was comparable to the ASEP/MD result. With this static (frozen) multipole solvent representation, we tried to locate a FS-MECI for the solute, and we could obtain it with a structure very similar to that found in the equilibrium solvent condition: a central bond length of 1.47 Å, dihedral of 90°, and BLA of 0.09 Å. Within this external perturbation, the FS-MECI is only 0.3 eV more stable than the FC point, while with ASEP/MD it is 1.58 eV lower at CASSCF level. These results point to the steric hindrance as the main reason of the loss of the CI in frozen solvent conditions and also evidence the importance that solvent relaxation has on the EQ-MECI stabilization (around 1.3 eV). Burghardt et al.,⁶ in a theoretical study of a simplified model with only one double bond and using a continuum method to represent the solvent, obtained a different result; they concluded that the loss of the CI in frozen solvent conditions was due to the destabilization of the ionic state by the reaction field generated by the solvent.

Turning to the fluorescence spectra, our data indicate that the low-frequency parts of the emission band could originate from different regions of the first excited surface, S_1 , of different electronic character, ionic and covalent, and we have thus a case of dual fluorescence. The high-frequency part is due to emission from the covalent minimum on S_1 , while the low-frequency part is due to emission from the ionic minimum also on S_1 . The relative energies of the two fluorescence maxima agree with the data reported by Zgrablić et al.²⁸ Given that the S_1/S_2 conical intersection is near the FC point, this double fluorescence appears mainly when the excitation energy is close or larger than the vertical excitation energy, a fact confirmed by the experiment. If the excitation energy corresponds to the 00 band (the point in which absorption and emission spectra intersect), only the ionic minimum seems to be populated.

Regarding the three decay components exhibited by the vibrationally cold S_1 fluorescence, Zgrablić et al.²⁸ conclude that they are due to the heterogeneity of the S_1 state in the sense that emission stems from several shallow potential surface minima. Our results have not permitted us to corroborate this conclusion. In gas phase, several ionic minima were found on S_1 , however in methanol solution, we found only one minimum. This minimum is very shallow; in fact, the free energy surface is very flat in the neighborhood of this minimum. So for instance, when the minima located in gas phase, Ion-1 and Ion-2, are solvated, their free energies differ only in 1.3 and 0.4 kcal/mol, respectively, from the ionic minimum found in methanol. The different relaxation times could correspond to molecules that follow slightly different paths on this flat free energy surface. However, given the different approximations employed in our calculations, we cannot exclude the existence of multiple minima on the S_1 surface. Multiexponential decay even with a single minimum has been proposed by Olivucci and co-workers when the minimum is separated from the CI by a shallow barrier, as is the case.⁵³ In this proposal the different decay times are due to different amounts of kinetic energy in the reactive torsion mode. Finally, an alternative mechanism, suggested by Hasson et al.,⁵⁴ could also explain the multiexponential decay. This can arise if the solvent fluctuations

modulate the barrier height leading to the reactive region on S_1 . In order to check the validity of this last approximation, the energy of the ionic minimum and of the barrier for some particular solvent configurations evenly distributed along the simulation was calculated. In both cases, the fluctuation measured as the standard deviation is 0.7 kcal/mol. The height of the barrier to be surmounted to access the CI region depends on the solvent configuration considered. The solvent not only modulates the barrier height but also changes the relative stability of covalent and ionic states at different points of the free energy surface. In Figure 4, for instance, at the geometry of the ionic minimum, the covalent state is 15.7 kcal/mol above the ionic state. This value has been obtained supposing the solvent in equilibrium with the charge distribution of the ionic state. If we recalculate the energies of the two excited states at the same geometry but with the solvent now in equilibrium with the charge distribution of the covalent state, then the stability order is reversed, i.e., the covalent state becomes more stable than the ionic state. The same occurs at the geometry of the covalent minimum, where the ionic state can become more stable than the covalent when one permits the solvent to equilibrate with the ionic state. In short, as said before, fluctuations in the solvent yield to the crossing of covalent and ionic states at both minimum geometries and at the FC point. The presence of a S_1/S_2 conical intersection modulated by solvent fluctuations could permit that part of the population be transferred from the ionic to the covalent excited state and back again and that could explain the various emission times found from the ionic excited state.

Given the different approximations of our calculations, including the use of CASSCF geometries and the neglect of vibrational contributions of the solute and the importance that solvent dynamics apparently has, there is a need for much more investigation before one can elucidate which of the proposals—several minima, different amounts of kinetic energy in the reactive torsion mode, or a distribution of barrier heights modulated by the solvent—is the correct one.

4. CONCLUSIONS

We have applied the ASEP/MD method to the study of solvent effects on the radiative and nonradiative decay of a model of retinal. In particular, we have focused our study on the first step in the *cis*–*trans* photoisomerization in methanol solution. Several ionic minima that differ in less than 1 kcal/mol have been located on S_1 in gas phase, whereas in methanol only one ionic minimum has been found. On the covalent excited surface several planar minima have also been found both in gas phase and in methanol. From a photochemical point of view, the most remarkable characteristic of the ionic excited state surface is the presence of the S_0/S_1 conical intersection. The optimized CI structures are similar in vacuum and in solution, with an important torsion of $\sim 90^\circ$ in the central double bond. In both phases, the evolution toward the conical intersection is practically a barrierless process. In gas phase, the de-excitation mainly proceeds nonradiatively through the conical intersection. In methanol solution, before the S_0/S_1 conical intersection can be reached, it needs a strong solvent reorganization, and the nonradiative route is slower; this permits the radiative decay.

We explain the structure of the emission band as a case of double fluorescence, where emission originates from two minima, ionic and covalent, on S_1 . After excitation, part of the population goes to the ionic excited-state minimum, from here some molecules

can decay radiatively giving origin to the low-frequency component of the fluorescence band, and others decay nonradiatively through the S_0/S_1 CI. We assign the high-frequency component of the fluorescence band to emission from the covalent minimum on S_1 . In gas phase only two states (ground and excited ionic) are involved in the photoisomerization. On the contrary, in methanol solution three states (ground, excited ionic, and excited covalent) are needed to describe the decay. Covalent and ionic states are degenerate in the neighborhood of the FC point, furthermore energy fluctuations modulated by the solvent cause the degeneration of these states at other geometries. This fact could permit the population transfer between the ionic and covalent states.

■ ASSOCIATED CONTENT

S Supporting Information. MS-CASPT2 results in methanol solution together with the CASSCF and CASPT2 energies for first three singlet states, the weights of the reference zeroth-order CASSCF wave function in the CASPT2 perturbation, the mixing coefficient of the most important state in the MS-CASPT2 treatment, and finally, the off-diagonal elements in the MS-CASPT2 symmetric Hamiltonian depending on the active space and the basis set. This material is available free of charge via the Internet at <http://pubs.acs.org>.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: auroram@unex.es.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENT

This work was supported by the CTQ2008-06224/BQU Project from the Ministerio de Ciencia e Innovación of Spain, cofinanced by the European Regional Development Fund (ERDF), and the PRI08A056 Project from the Consejería de Economía, Comercio e Innovación of the Junta de Extremadura. I.F.G. acknowledges the Junta de Extremadura and the European Social Fund for financial support. A.M.-L. acknowledges financial support from the Juan de la Cierva subprogramme of the Ministerio de Ciencia e Innovación of Spain. The authors also thank the Fundación Computación y Tecnologías Avanzadas de Extremadura (COMPUTAEX) for additional computational resources.

■ REFERENCES

- (1) Yarkony, D. R. In *Conical Intersections, Advanced Series in Physical Chemistry*; Domcke, W., Yarkony, D. R., Köppel, H., Eds.; World Scientific: Singapore, 2004; Chapter 2, No. 15, p 41.
- (2) Olivucci, M.; Sinicropi, A. In *Computational Photochemistry, Theoretical and Computational Chemistry*; Olivucci, M., Ed.; Elsevier: Amsterdam, The Netherlands, 2005; Chapter I, No. 16, p 1.
- (3) Yarkony, D. R. *J. Chem. Phys.* **1990**, *92*, 2457.
- (4) Atchity, G. J.; Xantheas, S. S.; Ruedenberg, K. *J. Chem. Phys.* **1991**, *95*, 1862.
- (5) (a) Reguero, M.; Olivucci, M.; Bernardi, F.; Robb, M. A. *J. Am. Chem. Soc.* **1994**, *116*, 2103. (b) Ragazos, N.; Robb, M. A.; Bernardi, F.; Olivucci, M. *Chem. Phys. Lett.* **1992**, *197*, 217.

- (6) Burghardt, I.; Cederbaum, L.; Hynes, J. T. *Faraday Discuss.* **2004**, *127*, 395.
- (7) Spezia, R.; Burghardt, I.; Hynes, J. T. *Mol. Phys.* **2006**, *104*, 903.
- (8) Santoro, F.; Barone, V.; Gustavsson, T.; Improta, R. *J. Am. Chem. Soc.* **2006**, *128*, 16312.
- (9) Yamazaki, S.; Kato, S. *J. Chem. Phys.* **2005**, *123*, 114510.
- (10) (a) Ten-no, S.; Hirata, F.; Kato, S. *J. Chem. Phys.* **1994**, *100*, 7443. (b) Sato, H.; Hirata, F.; Kato, S. *J. Chem. Phys.* **1996**, *105*, 1546.
- (11) Mori, T.; Nakano, K.; Kato, S. *J. Chem. Phys.* **2010**, *133*, 064107.
- (12) (a) Toniolo, A.; Granucci, G.; Martínez, T. J. *J. Phys. Chem. A* **2003**, *107*, 3822. (b) Toniolo, A.; Ben-Nun, M.; Martínez, T. J. *J. Phys. Chem. A* **2002**, *106*, 4679.
- (13) Garavelli, M.; Rugen, F.; Ogliano, F.; Bearpark, M. J.; Bernardi, F.; Olivucci, M.; Robb, M. A. *J. Comput. Chem.* **2003**, *24*, 1357.
- (14) Ciminelli, C.; Granucci, G.; Persico, M. *Chem.—Eur. J.* **2004**, *10*, 2327.
- (15) (a) Muñoz-Losa, A.; Fdez. Galván, I.; Martín, M. E.; Aguilar, M. A. *Chem. Phys. Lett.* **2007**, *443*, 76. (b) Muñoz-Losa, A.; Fdez. Galván, I.; Sánchez, M. L.; Martín, M. E.; Aguilar, M. A. *J. Phys. Chem. B* **2008**, *112*, 877.
- (16) Schoenlein, W. P.; Peteanu, L. A.; Mathies, R. A.; Shank, C. V. *Science* **1991**, *254*, 412.
- (17) Wang, Q.; Schoenlein, R. W.; Peteanu, L. A.; Mathies, R. A.; Shank, C. A. *Science* **1994**, *266*, 422.
- (18) (a) Garavelli, M.; Celani, P.; Bernardi, F.; Robb, M. A.; Olivucci, M. *J. Am. Chem. Soc.* **1997**, *119*, 6891. (b) Gonzalez-Luque, R.; Garavelli, M.; Bernardi, F.; Merchan, M.; Robb, M. A.; Olivucci, M. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 9379.
- (19) (a) Cembran, A.; Bernardi, F.; Olivucci, M.; Garavelli, M. *J. Am. Chem. Soc.* **2004**, *126*, 16018. (b) Cembran, A.; Bernardi, F.; Olivucci, M.; Garavelli, M. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6255.
- (20) Coto, P. B.; Strambi, A.; Olivucci, M. *Chem. Phys.* **2008**, *347*, 483.
- (21) Send, R.; Sundholm, D. *J. Phys. Chem. A* **2007**, *111*, 8766.
- (22) Warshel, A.; Chu, Z. T. *J. Phys. Chem. B* **2001**, *105*, 9857.
- (23) Finley, J.; Malmqvist, P.-Å.; Roos, B. O.; Serrano-Andrés, L. *Chem. Phys. Lett.* **1998**, *288*, 299.
- (24) Schoenlein, R. W.; Peteanu, L. A.; Mathies, R. A.; Shank, C. V. *Science* **1991**, *254*, 412.
- (25) Kandori, H.; Katsuta, Y.; Ito, M.; Sasabe, H. *J. Am. Chem. Soc.* **1995**, *117*, 2669.
- (26) Kandori, H.; Sasabe, H.; Nakanishi, K.; Yoshizawa, T.; Mizukami, T.; Shichida, Y. *J. Am. Chem. Soc.* **1996**, *118*, 1002.
- (27) (a) Polli, D.; Altoè, P.; Weingart, O.; Spillane, K. M.; Manzoni, C.; Brida, D.; Tomasello, G.; Orlandi, G.; Kukura, P.; Mathies, R. A.; Garavelli, M.; Cerullo, G. *Nature* **2010**, *467*, 440. (b) Frutos, L. M.; Andruniów, T.; Santoro, F.; Ferré, N.; Olivucci, M. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 7764.
- (28) Zgrablić, G.; Haacke, S.; Chergui, M. *J. Phys. Chem. B* **2009**, *113*, 4384.
- (29) Ishida, T.; Nanbu, S.; Nakamura, H. *J. Phys. Chem. A* **2009**, *113*, 4356.
- (30) (a) Fdez. Galván, I.; Sánchez, M. L.; Martín, M. E.; Olivares del Valle, F. J.; Aguilar, M. A. *Comput. Phys. Commun.* **2003**, *155*, 244. (b) Sánchez, M. L.; Martín, M. E.; Aguilar, M. A.; Olivares del Valle, F. J. *J. Comput. Chem.* **2000**, *21*, 705. (c) Martín, M. E.; Sánchez, M. L.; Olivares del Valle, F. J.; Aguilar, M. A. *J. Chem. Phys.* **2002**, *116*, 1613. (d) Sánchez, M. L.; Martín, M. E.; Fdez. Galván, I.; Olivares del Valle, F. J.; Aguilar, M. A. *J. Phys. Chem. B* **2002**, *106*, 4813.
- (31) Bearpark, M. J.; Robb, M. A.; Schlegel, H. B. *Chem. Phys. Lett.* **1994**, *223*, 269.
- (32) (a) Chandrasekhar, J.; Smith, S. F.; Jorgensen, W. L. *J. Am. Chem. Soc.* **1985**, *107*, 154. (b) Chandrasekhar, J.; Jorgensen, W. L. *J. Am. Chem. Soc.* **1985**, *107*, 2974. (c) Jorgensen, W. L. *Acc. Chem. Res.* **1989**, *22*, 184.
- (33) (a) Muñoz-Losa, A.; Fdez. Galván, I.; Martín, M. E.; Aguilar, M. A. *J. Phys. Chem. B* **2006**, *110*, 18064. (b) Muñoz-Losa, A.; Fdez. Galván, I.; Aguilar, M. A.; Martín, M. E. *J. Phys. Chem. B* **2008**, *112*, 8815.

- (34) Nielsen, I. B.; Lammich, L.; Andersen, L. H. *Phys. Rev. Lett.* **2006**, *96*, 018304.
- (35) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Zakrzewski, V. G.; Montgomery, J. A., Jr.; Stratmann, R. E.; Burant, J. C.; Dapprich, S.; Millam, J. M.; Daniels, A. D.; Kudin, K. N.; Strain, M. C.; Farkas, O.; Tomasi, J.; Barone, V.; Cossi, M.; Cammi, R.; Mennucci, B.; Pomelli, C.; Adamo, C.; Clifford, S.; Ochterski, J.; Petersson, G. A.; Ayala, P. Y.; Cui, Q.; Morokuma, K.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Cioslowski, J.; Ortiz, J. V.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Gomperts, R.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Gonzalez, C.; Challacombe, M.; Gill, P. M. W.; Johnson, B. G.; Chen, W.; Wong, M. W.; Andres, J. L.; Head-Gordon, M.; Replogle, E. S.; Pople, J. A. *Gaussian 98*, revision A11.3; Gaussian, Inc.: Pittsburgh, PA, 1998.
- (36) Refson, K. *Comput. Phys. Commun.* **2000**, *126*, 310.
- (37) Karlström, G.; Lindh, R.; Malmqvist, P.-Å.; Roos, B. O.; Ryde, U.; Veryazov, V.; Widmark, P.-O.; Cossi, M.; Schimmelpfennig, B.; Neogrady, P.; Seijo, L. *Comput. Mater. Sci.* **2003**, *28*, 222.
- (38) Ghigo, G.; Roos, B. O.; Malmqvist, P.-Å. *Chem. Phys. Lett.* **2004**, *396*, 142.
- (39) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Groud, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Cladwell, J. W.; Kollman, P. A. *J. Am. Chem. Soc.* **1995**, *117*, 5179.
- (40) Rajamani, R.; Gao, J. *J. Comput. Chem.* **2002**, *23*, 96.
- (41) Röhrig, U. F.; Guidoni, L.; Rothlisberger, U. *ChemPhysChem* **2005**, *6*, 1836.
- (42) Platz, P. E.; Moheler, J. H. *Biochemistry* **1975**, *14*, 2340.
- (43) Levine, B. G.; Martinez, T. J. *Annu. Rev. Phys. Chem.* **2007**, *58*, 613.
- (44) Valsson, O.; Filippi, C. *J. Chem. Theory Comput.* **2010**, *6*, 1275.
- (45) Garavelli, M.; Vreven, T.; Celani, P.; Bernardi, F.; Robb, M. A.; Olivucci, M. *J. Am. Chem. Soc.* **1998**, *120*, 1285.
- (46) Warshel, A. *Nature* **1976**, *260*, 679.
- (47) Schapiro, I.; Weingart, O.; Buss, V. *J. Am. Chem. Soc.* **2009**, *131*, 16.
- (48) Rostov, I. V.; Amos, R. D.; Kobayashi, R.; Scalmani, G.; Frisch, M. J. *J. Phys. Chem. B* **2010**, *114*, 5547.
- (49) Kochendoerfer, G. G.; Mathies, R. A. *J. Phys. Chem.* **1996**, *100*, 14526.
- (50) Fdez. Galván, I.; Sánchez, M. L.; Martín, M. E.; Olivares del Valle, F. J.; Aguilar, M. A. *J. Chem. Phys.* **2003**, *118*, 255.
- (51) Serrano-Andrés, L.; Merchán, M.; Lindh, R. *J. Chem. Phys.* **2005**, *122*, 104107.
- (52) Humphrey, W.; Dalke, A.; Schulten, K. *J. Mol. Graphics* **1996**, *14*, 33.
- (53) Olivucci, M.; Lami, A.; Santoro, F. *Angew. Chem., Int. Ed.* **2005**, *44*, 5118.
- (54) Hasson, K. C.; Gai, F.; Anfinrud, P. A. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 15124.

Chemiluminescence and Fluorescence States of a Small Model for Coelenteramide and Cypridina Oxyluciferin: A CASSCF/CASPT2 Study

Daniel Roca-Sanjuán,^{*,†} Mickael G. Delcey,[†] Isabelle Navizet,[‡] Nicolas Ferré,[§] Ya-Jun Liu,^{||} and Roland Lindh^{*,†}

[†]Department of Chemistry—Ångström, Theoretical Chemistry Programme, Uppsala University, P.O. Box 518, S-75120 Uppsala, Sweden

[‡]Molecular Science Institute School of Chemistry, University of the Witwatersrand, PO Wits Johannesburg 2050, South Africa

[§]Universités d'Aix-Marseille I, II, et III-CNRS UMR 6264: Laboratoire Chimie Provence, Equipe: Chimie Théorique Faculté de St-Jérôme, Case 521, 13397 Marseille Cedex 20, France

^{||}Key Laboratory of Theoretical and Computational Photochemistry, Ministry of Education, College of Chemistry, Beijing Normal University, Beijing 100875, China

S Supporting Information

ABSTRACT: Fluorescence and chemiluminescence phenomena are often confused in experimental and theoretical studies on the luminescent properties of chemical systems. To establish the patterns that distinguish both processes, the fluorescent and chemiluminescent states of 2-acetamido-3-methylpyrazine, which is a small model of the coelenterazine/coelenteramide and Cypridina luciferin/oxyluciferin bioluminescent systems, were characterized by using the complete active space second-order perturbation (CASPT2) method. Differences in geometries and electronic structures among the states responsible for light emission were found. On the basis of the findings, some recommendations for experimental studies on chemiluminescence are suggested, and more appropriate theoretical approaches are proposed.

INTRODUCTION

According to the International Union of Pure and Applied Chemistry (IUPAC) Gold Book,¹ chemiluminescence is the “emission of radiation resulting from a chemical reaction,” while fluorescence is the “luminescence which occurs essentially only during the irradiation of a substance by electromagnetic radiation.” Both phenomena can be described mechanistically by means of the scheme displayed in Figure 1. In the fluorescence process, the Franck–Condon structure is relaxed upon light absorption toward a minimum energy point in the excited state. From this *fluorescent state* (FS), the molecule emits radiation. The general picture for the chemiluminescence reaction mechanism, as established nowadays, is the following: (1) The reactants reach the transition state (TS) via a thermal movement, for example, in peroxy compounds, increasing mainly the OO bond distance. (2) A region of near-degeneracy or a conical intersection (CI) seam between the ground and excited state follows, promoting the system to the electronically excited hypersurface—the formation of a stable intermediate in this step is possible—and (3) the molecule decays from the *chemiluminescence state* (CS) to the ground state by means of light emission.^{2–5} Both fluorescent and chemiluminescent processes imply the emission of light from an excited state of the system. However, as we prove in this contribution, they correspond to different phenomena, and the excited state responsible for the emission of radiation are not necessarily the same.

Chemiluminescence and bioluminescence phenomena—the latter corresponding to the luminescence taking place in a living organism—have attracted much experimental and theoretical attention not only as a subject of academic interest in the field of

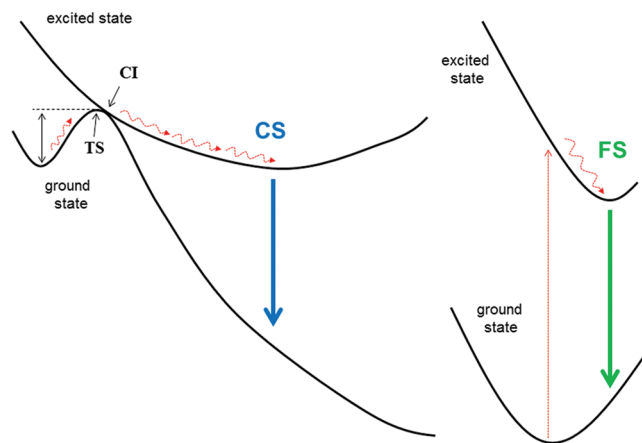


Figure 1. General scheme of the chemiluminescence (left) and fluorescence (right) processes. The chemiluminescence (CS) and fluorescence (FS) states are illustrated. The transition state (TS) and conical intersection (CI) points related to the former phenomenon are also shown.

biochemistry but also in the development of luminescence-based analytical techniques and the design of energy saving materials, where the chemical properties of chemiluminescent or bioluminescent systems are taken into account.^{6–8} Since the first appearance in 1968 of a monograph focused exclusively on the topic,⁹ many experimental results have been published, showing

Received: July 8, 2011

Published: November 02, 2011

a large list of chemiluminescent autoxidation and bioluminescent, enzyme-catalyzed, reactions (see for instance refs 10 and 11). A common pattern of some of these chemiluminescent reactions and almost all of the luminescent processes taking place in living organisms is the formation and decomposition of a molecule containing a peroxy intermediate, either 1,2-dioxetane or 1,2-dioxetanone (hereafter dioxetanone), resulting in the formation of a species in the excited state which emits light. Much effort has been focused on understanding the mechanism in small molecules containing these structures. Lindh and co-workers have been carrying out over the past five years a series of complete active space second-order perturbation (CASPT2) studies on the decomposition of 1,2-dioxetane and dioxetanone.^{2,3} A two-step mechanism was found in both cases. First, a biradical intermediate is formed through a TS, which is characterized by an OO stretching (dioxetanone) or a combination of OO stretching and OCCO torsion (1,2-dioxetane). The subsequent cleavage of the CC bond can bring the system to the excited state responsible for light emission. Meanwhile, differences at the molecular level were found in the study of a thiazole-substituted dioxetanone containing a π electron-donating group. This small model of the firefly luciferin molecule was employed to analyze the chemiluminescence reaction by means of the CASPT2 method.⁴ According to these results, and in agreement with other density functional theory (DFT) studies¹² with molecules containing the dioxetanone moiety coupled to an electron-donating group, a concerted charge transfer induced luminescence (CTIL) mechanism takes place. The electron-donating group, with a low ionization potential, partially delocalizes the π -electron toward the dioxetanone moiety, via the π system, decreasing the energy barrier of the reaction and enhancing the efficiency of the luminescence.

Reliable predictions of the whole chemiluminescent/bioluminescent mechanism require finding the point on the potential energy surface (PES) where the excited state can be populated efficiently and following the reactivity in this electronically excited PES. Modern computational chemistry offers several methods for describing electronically excited states of small to medium-size systems. However, the accurate prediction of excited electronic states in large-size systems is still a challenge, although many improvements are currently explored in the CASPT2 method to enable applications to large systems.¹³ Therefore, different approximations have been employed in the theoretical studies on chemiluminescence and bioluminescence. A commonly used strategy, which avoids following the photo-reactivity in the excited state, is to characterize the path toward the TS of the luminescent reaction before the nonadiabatic crossing (see Figure 1) by means of the DFT method and, next, make use of the Time-Dependent DFT (TD-DFT) and configuration interaction singles (CIS) methodologies to model the properties of the excited state around the region of the equilibrium structure for the product of the thermal decomposition reaction. Several imidazo[1,2-*a*]pyrazin-3(7*H*)-one (imidazopyrazinone) derivatives and dioxetanone-containing molecules with a substituent of low oxidation potential have been studied by using this approach.^{12,14} In other cases, charge-transfer excitations over the π -conjugated system of the reaction product are used to extract conclusions about the concerted CTIL mechanism, although the CTIL is the excitation to the antibonding- σ orbital of the OO bond in the dioxetanone moiety, which facilitates the decomposition of the molecule by lowering the activation energy required for bond breaking.¹⁵

On the other hand, several experiments were designed in the past to study the chemiluminescence properties of different molecular systems by irradiating the spent solution after chemiluminescence and measuring the spectrum of the radiation emitted, or directly by studying the fluorescent properties of the final products, such as distinct phenolate anions of coelenteramide analogues.^{16–18} Other experimental determinations of the light-emitting species are based on matching the chemiluminescence spectrum with the fluorescence. Examples can be found in the studies on the coelenterazine/coelenteramide^{19,20} system, or the luciferin/oxy-luciferin molecules from the *Cypridina* (Vargula)²¹ and firefly²² organisms. Indeed, the bioluminescence and fluorescence spectra show noticeable differences, as was proved by Belogurova et al. through the deconvolution of the bioluminescence spectra of photoproteins from marine coelenterates and the photoluminescence spectra of the bioluminescent reaction products.²³ These differences are commonly assigned to different forms of the product [nonionized, amide, ion-pair proton transfer, phenolate, pyrazine-N(4) anion forms].

Care must be taken with both mentioned experimental and theoretical procedures, since the region reached on the excited-state surface need not correspond in general to the structure responsible for light emission in the chemiluminescence or bioluminescence processes. To establish the molecular basis of the differences between both luminescent phenomena, we employed 2-acetamido-3-methylpyrazine. This molecule is similar to the product of decomposition of other imidazopyrazinone compounds reported in the literature¹⁴ and corresponds to a small model of coelenteramide and *Cypridina* oxy-luciferin, which are the light-emitting species present in some marine bioluminescent organisms, such as the hydromedusa *Aequorea*²⁴ and the hydroid *Obelia*,²⁵ in the first case, and the ostracod *Vargula hilgendorffii*,²⁶ in the latter (see Figure 2). The bioluminescent process in these organisms involves the protein systems aequorin, obelin, and *Cypridina* luciferase, respectively, which are formed by the apoprotein, coelenterazine (in the two former) or *Cypridina* luciferin (in the latter), and molecular oxygen. These complexes emit blue light through an intramolecular reaction, decomposing into apoprotein, coelenteramide (in aequorin and obelin) or *Cypridina* oxy-luciferin (in *Cypridina* luciferase), and CO₂.²⁴ The present study will contribute to the understanding of the bioluminescent properties of the coelenterazine/coelenteramide and *Cypridina* luciferin/oxy-luciferin systems. Since there is a controversy over the protonation state of the light-emitting species,²⁷ both anion and neutral forms were considered. The characterization of the CS and FS states was carried out by means of the CASPT2 method, finding relevant differences at the molecular level.

■ METHODS AND COMPUTATIONAL DETAILS

The double- ζ ANO-S²⁸ basis set with a contraction scheme [3s2p1d] for carbon, oxygen, and nitrogen atoms and [2s1p] for hydrogen (i.e., the ANO-S-VDZP basis set) was used throughout. Geometry optimizations, minimum energy reaction paths (MEPs), and the determination of surface crossings were carried out at the complete active space self-consistent field (CASSCF) level of theory with no symmetry restriction (*C*₁ symmetry), and the energies of the single points were corrected including the second-order perturbation treatment (CASPT2).^{13,29} In some cases, geometry optimizations were performed at the CASPT2 level, by using numerical gradients, to analyze the effect of the

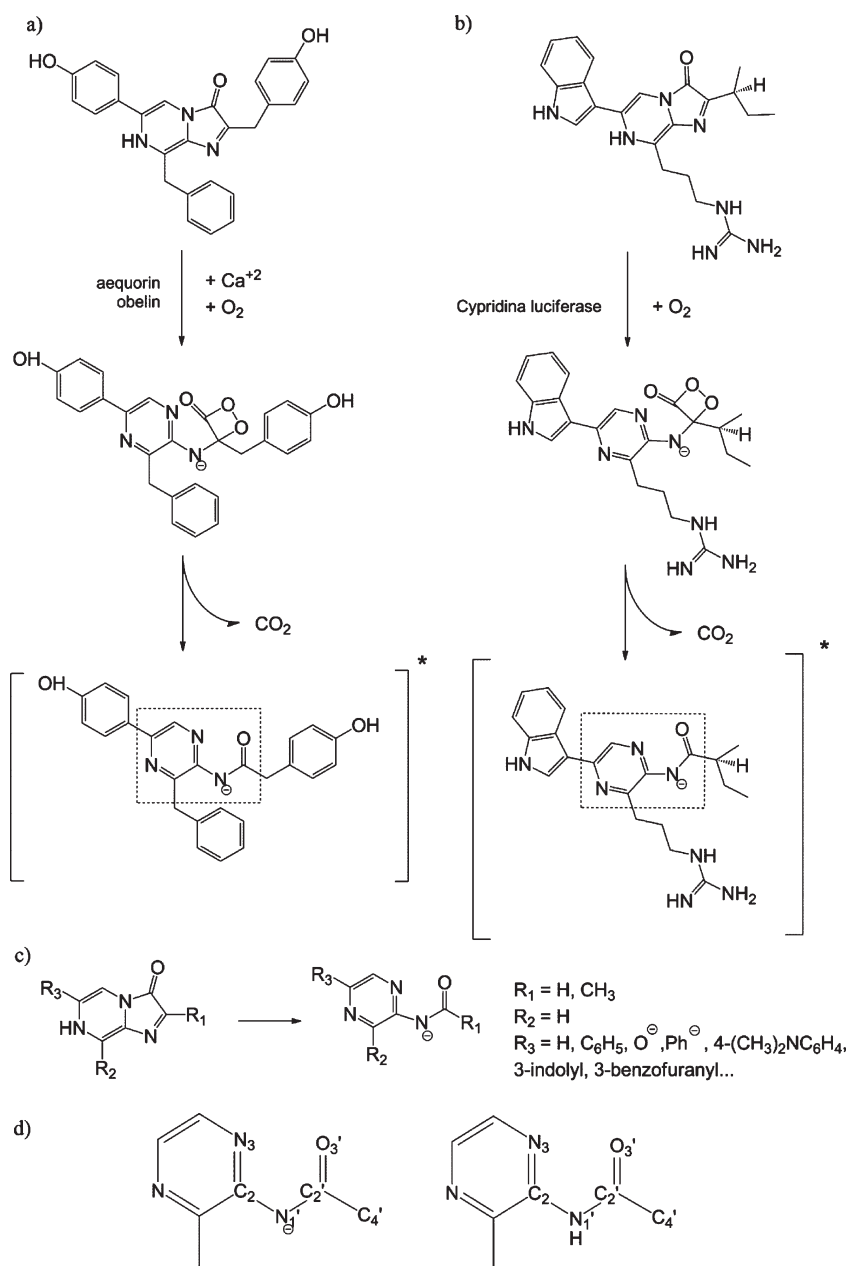


Figure 2. Bioluminescence reactions in the coelenterazine/coelenteramide (a) and *Cypridina* luciferin/oxy luciferin (b) systems. Two steps are represented: (1) The peroxy intermediate is produced from the reaction between coelenterazine (a) or *Cypridina* luciferin (b) and molecular oxygen in the aequorin/obelin (a) or *Cypridina* luciferase (b) proteins; (2) the peroxy intermediate decomposes, creating coelenteramide (a) or *Cypridina* oxy luciferin (b) in the excited state. (c) Distinct imidazo[1,2-*a*]pyrazin-3(7*H*)-one (imidazopyrazinone) derivatives employed in the experimental and theoretical studies on the bioluminescence and chemiluminescence of coelenterazine/coelenteramide and *Cypridina* luciferin/oxy luciferin (left) and the product of the chemical reaction (right). (d) Chemical structure of the anionic (left) and neutral (right) small models studied in the present contribution with the labels used in the text.

dynamical correlation in the geometries and the reliability of the equilibrium structures that were obtained with the noncorrelated CASSCF method. MEPs were built as steepest descendent paths in which each step requires the minimization of the PES on a hyperspherical cross-section of the PES centered on the initial geometry and characterized by a predefined radius.³⁰ Mass-weighted coordinates were used; therefore the MEP coordinate corresponds to the so-called intrinsic reaction coordinate (IRC), measured in atomic units, that is, $\text{bohr} \cdot \text{amu}^{1/2}$. The CI was computed by using the restricted Lagrange multipliers technique

as included in the MOLCAS-7.6 package,³¹ in which the lowest-energy point was obtained under the restriction of degeneracy between the two considered states.³⁰ In all of the CASPT2 calculations, the core orbitals of non-hydrogen atoms were not correlated, an imaginary level-shift of 0.2 au was turned on a priori to minimize weakly interacting intruder states, and the nonstandard IPEA modification of the zeroth-order Hamiltonian with a value of 0.00 au was employed. The ANO-S-VDZP basis set and nonstandard IPEA are commonly used in studies on photoreactivity (see, for instance, González-Ramírez et al.³² and

Table 1. Vertical Absorption Energies (in eV) and Oscillator Strengths (f) for the Anionic System of 2-Acetamido-3-methylpyrazine

transition	ΔE	f
$\pi_{\text{HOMO}} \rightarrow \pi_{\text{LUMO}}^*$	3.25	0.113
$\pi_{\text{HOMO}} \rightarrow \pi_{\text{LUMO}+3}^*$	4.03	0.436
$\pi \rightarrow \pi^*$	5.03	0.068
$\pi_{\text{HOMO}} \rightarrow \pi_{\text{CO}}^*$ (CTIL)	7.19	0.056

Table 2. Vertical Absorption Energies (in eV) and Oscillator Strengths (f) for the Neutral System of 2-Acetamido-3-methylpyrazine

transition	ΔE	f
$\pi_{\text{HOMO}} \rightarrow \pi_{\text{LUMO}}^*$	4.27	0.066
$\pi_{\text{HOMO}} \rightarrow \pi_{\text{LUMO}+3}^*$	5.14	0.222
$\pi \rightarrow \pi^*$	6.04	0.061
$\pi_{\text{HOMO}} \rightarrow \pi_{\text{CO}}^*$ (CTIL)	7.24	0.152

references therein). Further test calculations were carried out with the ANO-RCC type of basis set or the standard IPEA (0.25 au), as used in previous studies on chemiluminescence,^{3,4} and they do not change the conclusions.

Two sets of active spaces were used in the CASSCF models (see Figures S1 and S2, Supporting Information). In order to characterize the chemiluminescent and fluorescent properties of the molecules, an optimal candidate for the calculations is an active space of 10 electrons distributed in nine orbitals [denoted as CAS(10-in-9)]. All of the π orbitals of the six-member-ring, $\text{N}_{1'}$ atom, and carbonyl group are considered in this case. It has been shown elsewhere that a lone pair of the oxygen atom may be relevant to the chemiluminescent reactivity.^{3,4} Therefore, an active space including this orbital [CAS(12-in-10)] was also employed for the calculations related to the chemiluminescent state.

All of the computations were carried out by using the MOLCAS 7.6 quantum-chemical software.³¹

RESULTS AND DISCUSSION

Fluorescence State (FS). The CASPT2 vertical spectra of the anionic and neutral molecules computed at the minimum structure of the ground state [hereafter (S_0)_{min}] are compiled in Tables 1 and 2, respectively, and the shapes of the natural orbitals (NOs) involved in the most relevant excitations are depicted in Figure 3. The lowest-lying excited states (S_1 and S_2) are related in both molecules to the same transitions between NOs delocalized over the π -aromatic system. While S_1 can be characterized by an excitation from the highest occupied to the lowest unoccupied molecular orbitals (HOMO and LUMO, respectively) from the previous self-consistent field (SCF) calculation, the HOMO and LUMO+3 orbitals are involved in S_2 (see Figure 3 and Tables 1 and 2). S_2 has the highest oscillator strength (f), and therefore will be the most populated state upon irradiation. Lower energies are required to excite the anionic system. In particular, the brightest state of the anion requires 4.03 eV, and the lowest-energy state appears at 3.25 eV, while the corresponding transitions in the neutral system are located more than 1 eV higher in energy at 5.14 and 4.27 eV, respectively. Direct comparisons with the experimental absorption

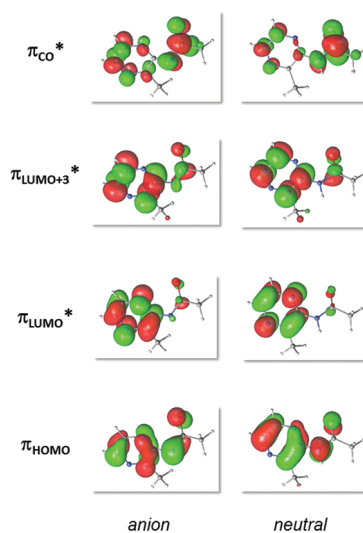


Figure 3. Selected natural orbitals (NOs) to analyze the absorption and emission properties of the anionic and neutral systems of 2-acetamido-3-methylpyrazine (see text). The complete sets of NOs present in the active space of the CASSCF/CASPT2 calculations can be found in Figures S1 and S2 (Supporting Information).

spectra of coelenteramide are difficult since the transitions will be mainly affected by the phenolate substituent of the pyrazine ring (see Figure 2). Nevertheless, two peaks are also found in different coelenteramide analogues studied by Shimomura and Teranishi.¹⁸ As explained above and proved in several studies,^{4,12} efficient chemiluminescent systems are governed by a charge-transfer excitation from an electron-donating part of the molecule to the dioxetanone moiety (concerted CTIL mechanism). In the systems studied here at the Franck–Condon geometry, none of the lowest excited states are characterized by a CTIL excitation, even in the anion species in which the $\text{N}_{1'}$ atom is charged negatively, and therefore the system presents a lower oxidation potential.¹² To find a state with charge-transfer character to the carbonyl group, more than 10 roots are needed in the State Averaged CASSCF (SA-CASSCF) wave function. These transitions correspond to excitations from the SCF HOMO to the π^* orbital located mainly in the CO group (see Figure 3) and require energies larger than 7 eV (see Tables 1 and 2). According to these results, the states with CTIL character, which are involved in the chemiluminescence phenomenon, are not accessible at the optimized geometry of the ground state of 2-acetamido-3-methylpyrazine anion and neutral molecules.

The minima of the S_1 and S_2 states for the anion and neutral molecules were subsequently determined by means of the CASSCF/CASPT2 method to analyze the fluorescent properties of these systems. A minimum point in the PESs of the S_2 state is found relatively close in energy to the S_1 surface (0.29 eV higher), with a vertical emission to the ground state (VE) of 2.39 eV and a f of 0.001, while the S_1 geometry optimization results in a structure with a VE of 2.13 eV and a larger f (0.043), see Table 3. In the neutral molecule, the equilibrium structures have higher VEs, similarly to the results found in the absorption computations. In particular, S_2 and S_1 present a VE of 2.75 and 4.07 eV, respectively, with a higher probability of fluorescence from the latter (the f values are 0.001 and 0.046, respectively). The theoretical description of the fluorescence process in the anionic

Table 3. Vertical Emission Energies (in eV) and Oscillator Strengths (f) for the Anionic System of 2-Acetamido-3-Methylpyrazine in the Fluorescence (FS) and Chemiluminescence (CS) States^a

		ΔE	f
fluorescence	$(S_1)_{\min}$	2.11	0.040
	$(S_2)_{\min 1}$	3.59	0.261
	$(S_2)_{\min 2}$	2.21	0.001
chemiluminescence	CS_{10in9}	2.59	0.006
	CS_{12in10}	1.72	0.002
	CS_{caspt2}	1.60	0.139

^a Fluorescent minima computed at the CASPT2//CASSCF(10-in-9) level of theory and chemiluminescence states obtained with different CASSCF/CASPT2 approaches (see text).

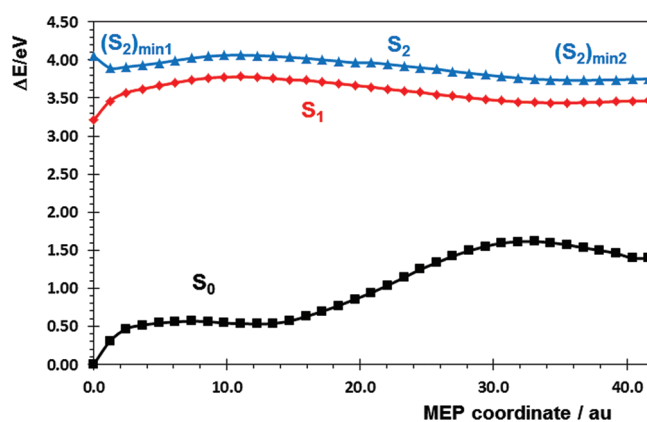


Figure 4. Low-lying singlet states of the 2-acetamido-3-methylpyrazine anion computed at the CASPT2//CASSCF level along the minimum energy path (MEP) of the S_2 state from the ground-state minimum. The S_2 -MEP ends in the equilibrium geometry of $(S_2)_{\min 2}$.

molecule was later improved by using the photochemical reaction path approach³³ to characterize with accuracy the emissive state of the molecule. The anion form of 2-acetamido-3-methylpyrazine is expected to be more relevant in the chemiluminescence, as the system has a better electron-donating component, and therefore a more exhaustive analysis is necessary. A MEP from the brightest state, S_2 , was performed (S_2 -MEP); Figure 4 displays the CASPT2 energies for the low-lying S_0 , S_1 , and S_2 states along the S_2 -MEP. The CASSCF S_2 -MEP (see Figure S3, Supporting Information) ends up in the region of the previously optimized minimum on the S_2 surface, and two minima are found when the dynamical correlation is included through the CASPT2 [hereafter, $(S_2)_{\min 1}$ and $(S_2)_{\min 2}$], as can be seen in Figure 4. The structure $(S_2)_{\min 1}$, which is found close to the ground-state minimum geometry, has a VE of 3.59 eV and a high probability of fluorescence ($f = 0.261$). This minimum is however very shallow, and from the Franck–Condon structure the system can surmount the energy barrier of 0.16 eV and reach the second minimum $(S_2)_{\min 2}$. The stability of this structure, $(S_2)_{\min 2}$, is also relatively small, since the S_1 surface is only 0.29 eV lower and a CI between both states is found energetically close to $(S_2)_{\min 2}$, in particular, only 0.16 eV above this minimum. From this CI, a MEP was computed on the PES of the S_1 state (S_1 -MEP) to study the evolution of the system after reaching the crossing point between both S_1 and S_2 surfaces (see Figure 5 and

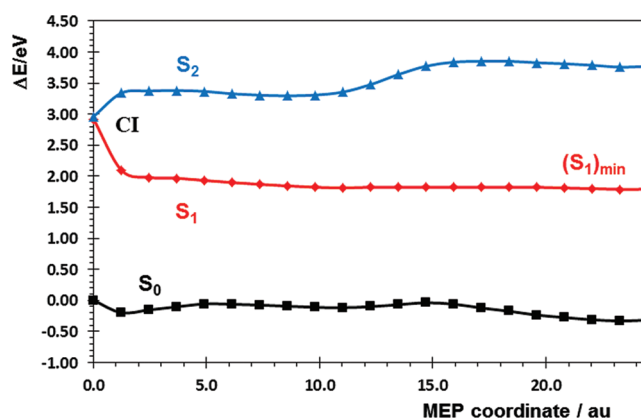


Figure 5. Low-lying singlet states of the 2-acetamido-3-methylpyrazine anion computed at the CASPT2//CASSCF level along the minimum energy path (MEP) of the S_1 state from the conical intersection between the S_2 and S_1 states (CI). The S_1 -MEP ends in the equilibrium geometry of $(S_1)_{\min}$.

Figure S4, Supporting Information). The S_1 minimum previously optimized at the CASSCF level [hereafter, $(S_1)_{\min}$] is reached at the end of the S_1 -MEP. As can be seen in Figure 4, S_1 and S_2 states are close in energy (<0.3 eV) over a large region of the MEP, which implies a high probability of hopping between the S_1 and S_2 surfaces. In order to investigate alternative photochemical paths that may occur and therefore to further characterize the fluorescent properties of the 2-acetamido-3-methylpyrazine anion, several geometry optimizations and a MEP were computed on the S_1 surface from different points of the S_2 -MEP. According to the computations of the f magnitude, the optical interaction between the S_2 and S_1 states decreases toward the CI. Thus, the MEP was performed from the equilibrium geometry $(S_2)_{\min 1}$ (see Figures S5 and S6, Supporting Information). All of these computations arrive at the same point, $(S_1)_{\min}$, suggesting that other structures are not relevant in the fluorescence of the molecule. In addition, since S_1 can also be populated after irradiation (see the value for f in Table 1), the evolution of this state from the Franck–Condon point was also characterized by means of a MEP (see Figures S7 and S8, Supporting Information). Again, the structure of $(S_1)_{\min}$ is reached at the end of the computation.

Figure 6 displays the energy profile of the photochemical paths analyzed. All three minima found will contribute in some extent to the fluorescence of the system. $(S_2)_{\min 1}$, even with a large f , is a shallow minimum, and a large amount of the molecules UV-irradiated will possess enough energy to access the second minimum and the CI and continue deactivating the excess energy toward the $(S_1)_{\min}$. Alternative paths derived from the hopping from the S_2 to S_1 surfaces along the region of close energy drive the system to the same emissive state, $(S_1)_{\min}$. From this structure, there are no accessible crossings with the ground state, and the system will fluoresce. The experimental emission spectrum of the amide anion form of coelenteramide shows only one fluorescence peak at 2.70–2.85 eV.¹⁸ Probably, it is related to the $(S_1)_{\min}$ structure obtained in the present study, although, as explained above, comparisons of the transition energies between coelenteramide and 2-acetamido-3-methylpyrazine are not straightforward.

The analysis of the electronic configuration and the geometrical parameters that are relevant to fluorescence is crucial for

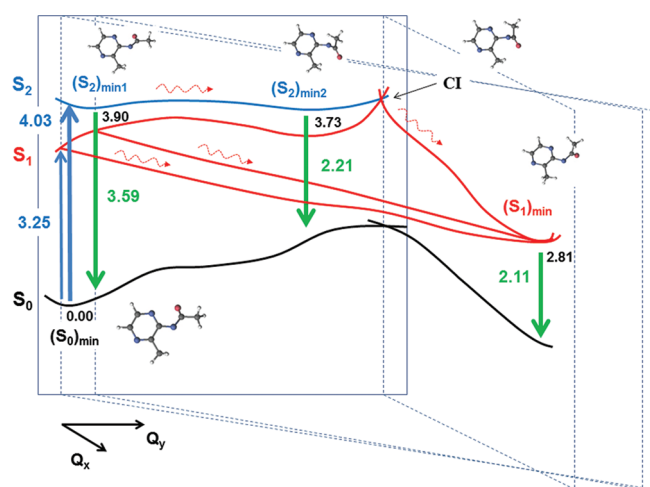


Figure 6. Scheme of the main decay paths on the singlet manifold of the 2-acetamido-3-methylpyrazine anion. Structures for the relevant points in the fluorescence process are shown. The numerical values (in eV) correspond to the vertical absorption (blue) and emission (green) energies computed at the CASPT2//CASSCF level. The Q_y coordinate is mainly related to the rotation of the acetamido group with respect to the methylpyrazine part around the $C_2N_1C_2N_3$ dihedral angle and to the puckering of the six-member ring, whereas Q_x is associated with the orientation of the CO and methyl groups relative to the rest of the molecule.

establishing the fluorescent patterns and distinguishing them from chemiluminescence. Excitations between NOs localized in the six-member ring characterize the electronic changes that govern the photophysics of the systems (see Figure 3). All of the optimized structures and points along the MEPs have this nature. Charge-transfer transitions to the carbonyl group are very high in energy, as proved in the computation of the absorption spectrum. To show the structural properties of the FS, the CO bond distance and the $O_3C_2N_1C_4'$ and $C_2N_1C_2N_3$ dihedral angles were selected (see Figure 2). The first angle shows out-of-plane distortions of the carbonyl group with respect to the adjacent N_1' and C_4' atoms. Meanwhile, the latter is related to the rotation of the acetamido group with respect to the plane defined by the methylpyrazine moieties. The values obtained for all of the relevant points in the fluorescent study of the 2-acetamido-3-methylpyrazine anionic and neutral systems are compiled in Table 4. The CO bond length for all minima of the anion [(S_1)_{min}, (S_2)_{min1}, and (S_2)_{min2}] are in the range 1.20–1.23 Å. Similar values are obtained in the geometry optimizations of S_2 and S_1 in the neutral species, in particular, 1.20 Å for both minima. These CO bond distances are close to those obtained for the ground-state minimum geometry and correspond to a double bond CO. Regarding the dihedral angles, different values are obtained for $C_2N_1C_2N_3$. As can be seen in Figure 6 and Table 4, this angle increases along the S_2 -MEP and reaches the largest value at the CI point where the acetamido group becomes perpendicular to the methylpyrazine moiety. From the CI and along the S_1 -MEP, the dihedral angle decreases and ends at a value of 4° in the (S_1)_{min} minimum structure. The behavior of the $O_3C_2N_1C_4'$ dihedral angle is different with respect to $C_2N_1C_2N_3$. The former is kept practically unchanged during the fluorescent process, with a value that differs by only 5° from a planar conformation of the carbonyl. In the neutral form, out-of-plane distortions are not obtained for neither the S_2 nor S_1

Table 4. Computed Structural Parameters for the Relevant Geometries in the Absorption and Emission Phenomena of 2-Acetamido-3-methylpyrazine^a

		CO	$O_3C_2N_1C_4'$	$C_2N_1C_2N_3$
literature ^b	Conf1	1.20	−178	−5
	Conf2	1.20	174	173
	Conf3	1.21	179	2
	Conf4	1.20	176	−170
absorption	(S_0) _{min}	1.25	−177	36
	(S_1) _{min}	1.20	−173	4
fluorescence	(S_2) _{min1}	1.23	178	52
	(S_2) _{min2}	1.20	−175	92
	CI	1.20	−176	95
	(S_1) _{min}	1.20	−173	4
chemiluminescence	CS _{10in9}	1.28	−132	3
	CS _{12in10}	1.40	−125	2
	CS _{caspt2}	1.33	−127	20

^aThe reported values for the emissive states of different conformers related to 2-acetamidopyrazine are also shown. Bond lengths are in Å and dihedral angles in degrees. ^bRelated geometrical parameters for the lowest excited state optimized structure of four conformers of the 2-methanamidopyrazine molecule computed at the CIS/6-31+G(d) level of theory. Data taken from ref 14.

minima. The findings for the CO bond length and $O_3C_2N_1C_4'$ dihedral angle show that the fluorescent events are characterized by a double-bond CO and an sp^2 hybridization of the C_2' atom, which is in agreement with the fact that the electronic excitations involved do not change markedly the π -bonding character of the carbonyl group. The structure of the fluorescence state is therefore far from the geometry of the peroxy intermediate which decomposes to the species responsible for the chemiluminescence, with a single-bond and sp^3 hybridization of the carbonyl moiety (see Figure 2).

Chemiluminescence State (CS). The CS is the point on the excited-state PESs reached after the decomposition of the dioxetane (or peroxy) precursor and is responsible for light emission (see Figure 2). As explained in the Introduction, it is common in the theoretical studies on chemiluminescence to start with the ground-state optimized structure of the reaction product in order to find the CS. This strategy is dangerous, since the molecule reaches the excited state at the region of energy crossing (CI) between the PESs of the ground and excited states in the vicinity of the TS related to the fragmentation reaction (see Figure 1), and the molecular structures at these points are closer to the peroxy intermediate than to the ground-state minimum of the reaction product, as has been proved in CASSCF/CASPT2 studies on the decomposition of 1,2-dioxetane and dioxetane,^{2,3} as well as DFT investigations with coelenterazine/coelenteramide models.¹⁴ From the CI point close to the TS, the photoreactivity continues in the excited state toward the CS. Therefore, a more appropriate strategy to find the CS might be to consider the peroxy molecule, which is the intermediate in the coelenterazine/coelenteramide and *Cypridina* luciferin/oxy-luciferin reactions (see Figure 2), as the starting structure in the study.

The ground state of the peroxy intermediate for the anionic and neutral small models of coelenteramide and *Cypridina* oxy-luciferin was optimized at the CASSCF level of theory. At these structures, the C_2' atom clearly shows sp^3 hybridization (the dihedral angles $O_3C_2N_1C_4'$ are around -125° for both systems)

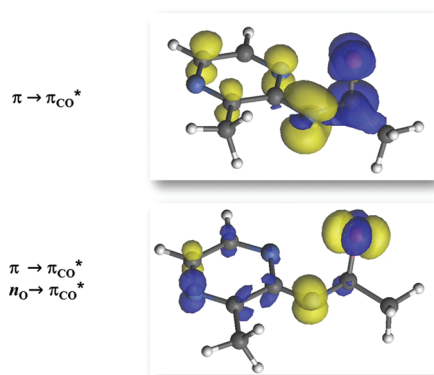


Figure 7. Electron-density difference between the ground and charge-transfer-induced-luminescence (CTIL) states of the peroxy-like form calculated with the CASSCF method and two active spaces: all π orbitals [CASSCF(10-in-9), top] and the π system plus the lone pair orbital n_{O} in the oxygen atom of the carbonyl group [CASSCF(12-in-10), bottom].

and the CO bond lengths are 1.53 and 1.47 Å, respectively. After the removal of the CO₂ moiety, resulting in 2-acetamido-3-methylpyrazine with a structure that will be named “peroxy-like form”, the lowest vertical transitions were calculated by means of the CASSCF/CASPT2 method and using only the π MOs in the active space. In contrast to the vertical electronic transitions at the ground-state minimum, the $\pi_{\text{HOMO}} \rightarrow \pi_{\text{CO}^*}$ (see Figure 3) excitation is present here in the lowest excited state, with a vertical absorption energy of 2.36 eV. The electron-density difference between the lowest excited and ground states is shown in Figure 7. This electronic transition corresponds to the CTIL state, which is characterized by the charge transfer from the π -aromatic system to the CO moiety and is responsible for enhancing the chemiluminescence efficiency.^{4,12} None of the lowest excited states of the neutral form with the peroxy-like structure has CTIL character. Thus, the presence of the deprotonated nitrogen (anion system) seems to be necessary in the molecule studied to promote the concerted CTIL mechanism for chemiluminescence. The charged nitrogen activates the electron donation from the π system by lowering the ionization potential of the HOMO.

In order to explore the properties of the PESs in the surroundings of the peroxy-like form, in the crossing region between the surfaces of the ground and excited states, and along the path toward the decomposition product, the electronic structure of the lowest lying excited state and the energy difference between the CTIL ($\pi_{\text{HOMO}} \rightarrow \pi_{\text{CO}^*}$) and $\pi_{\text{HOMO}} \rightarrow \pi_{\text{LUMO}^*}$ excited states and between the CTIL and the ground states were computed for several structures characterized by different C₂O₃ bond lengths and O₃C₂N₁C₄ dihedral angles, and at the structures obtained in CI calculations. The values for these geometrical parameters together with the results obtained in the analysis are compiled in Table 5. Shorter CO bond lengths with respect to the standard distance for the peroxy intermediate (1.54 Å), reaching a value as low as 1.20 Å, do not result in changes in the electronic nature of the lowest lying excited state; it maintains a CTIL character. The $\pi_{\text{HOMO}} \rightarrow \pi_{\text{LUMO}^*}$ excited state becomes progressively closer in energy to the CTIL state, with a minimum energy difference of 0.51 eV for the geometries considered here. Changes in the O₃C₂N₁C₄ dihedral angle result in much smaller energy separation and ultimately a change

Table 5. Electronic Transition Character of the Lowest Lying Excited State and Vertical CASPT2 Energies (in eV) between the $\pi_{\text{HOMO}} \rightarrow \pi_{\text{CO}^*}$ and $\pi_{\text{HOMO}} \rightarrow \pi_{\text{LUMO}^*}$ States [$\Delta E(S_2-S_1)$] and between the Lowest Lying Excited and Ground States [$\Delta E(S_1-S_0)$] at Different Geometries (see text)^a

	C ₂ O ₃	O ₃ C ₂ N ₁ C ₄	S ₁ nature	ΔE (S ₂ -S ₁)	ΔE (S ₁ -S ₀)
bond	1.30	-125	$\pi_{\text{HOMO}} \rightarrow \pi_{\text{CO}^*}$	1.03	1.65
	1.28	-125	$\pi_{\text{HOMO}} \rightarrow \pi_{\text{CO}^*}$	0.93	1.74
	1.26	-125	$\pi_{\text{HOMO}} \rightarrow \pi_{\text{CO}^*}$	0.83	1.84
	1.24	-125	$\pi_{\text{HOMO}} \rightarrow \pi_{\text{CO}^*}$	0.72	1.95
	1.22	-125	$\pi_{\text{HOMO}} \rightarrow \pi_{\text{CO}^*}$	0.62	2.05
	1.20	-125	$\pi_{\text{HOMO}} \rightarrow \pi_{\text{CO}^*}$	0.51	2.16
dihedral	1.30	-135	$\pi_{\text{HOMO}} \rightarrow \pi_{\text{CO}^*}$	0.88	1.85
	1.30	-143	$\pi_{\text{HOMO}} \rightarrow \pi_{\text{CO}^*}$	0.61	2.26
	1.30	-150	$\pi_{\text{HOMO}} \rightarrow \pi_{\text{CO}^*}$	0.57	2.69
	1.30	-158	$\pi_{\text{HOMO}} \rightarrow \pi_{\text{LUMO}^*}$	2.14	5.00
CASSCF CI	1.33	-121	$\pi_{\text{HOMO}} \rightarrow \pi_{\text{CO}^*}$	2.19	0.44
CASPT2 CI	1.32	-119	$\pi_{\text{HOMO}} \rightarrow \pi_{\text{CO}^*}$	2.47	0.02

^aBond lengths and dihedral angles are in Ångstroms and degrees, respectively.

of order (see Table 5). Hence, it is only when the molecule approaches the geometry of the decomposition product [(S₀)_{min}]—far from the region of the TS and the CI seam—that the electronic structure of the S₁ state becomes identical to the FS. This region presents a large energy difference between S₁ and the ground state, which indicates that the system is entering the region of the PESs governed by the fluorescence phenomenon.

CI computations at the CASSCF level of theory and including the π system result in a shorter CO bond length (a difference of 0.2 Å) and a similar O₃C₂N₁C₄ dihedral angle with respect to the peroxy-like form. Thus, the system still presents sp³ hybridization and single-bond character of the carbonyl group at this point. The electronic nature of the excited state also corresponds to the CTIL excitation. As can be seen in Table 5, strong differential correlation effects characterize the region of degeneracy: the CASPT2 energy difference between the ground and lowest excited states is 0.44 eV at the CI obtained with the CASSCF method. The CASPT2 CI was then computed by means of using the CASSCF gradients (see ref 34 for more details about this computational strategy). The molecule in the region of degeneracy at the CASPT2 level shows a similar geometry to the CASSCF CI point (differences are mainly related to rotations of the methyl group attached to the carbonyl moiety) and the same type of excitations between MOs for the lowest excited state. Both CI crossing points have relative energies of 3.51 and 3.36 eV with respect to the (S₀)_{min} equilibrium structure. These energy values must be considered only as an estimation of the CI, since, in contrast to the FS and CS, the molecule in the region of PES crossing is expected to possess the CO₂ part of the peroxy intermediate relatively close to the 2-acetamido-3-methylpyrazine moiety (see, for instance, refs 4 and 14), and therefore the energetic profile might be affected by this leaving group. In order to properly describe this part of the chemiluminescence reaction path, the CO₂ group and relevant MOs in this moiety should be included in the calculations, as was done in previous studies in 1,2-dioxetane,² 1,2-dioxetanone,³ and thiazole-substituted dioxetanone molecules.⁴ This is out of the

scope of the present contribution, which is focused on the FS and CS states. A relevant conclusion can still be obtained from the analysis performed here, which supports the use of the peroxy-like form as a starting structure to find the CS: the geometries at the decomposition reaction TS, in its vicinity, and in the region of PES crossing share in common a CTIL excited state and similar structural parameters for the CO moiety.

Subsequent geometry optimizations of the CTIL state of the peroxy-like form in the anion were carried out by using different CASSCF/CASPT2 strategies to determine the chemiluminescence properties and analyze different methodological parameters. First, a small active space (only π orbitals) was used to optimize the structure at the CASSCF level. In the second step, two modifications of the methodology were considered: (1) The lone pair n_{O} of the oxygen $\text{O}_{3'}$ was added since this orbital was shown in previous studies to contribute to the electronic transitions related to the CS, and (2) the dynamical correlation was used to optimize the geometry through the CASPT2 method by using numerical gradients. The results obtained for the VE energies and the selected geometrical parameters to monitor the differences between chemiluminescence and fluorescence are compiled in Tables 3 and 4. The multiconfigurational character of the CS becomes stronger when the n_{O} orbital is included in the active space. The CTIL and a new excitation from the n_{O} to the π_{CO}^* orbitals have large contributions in the electronic structure of the CS (see Figure 7). This results in, as a consequence, the elongation of the CO bond length (0.12 Å larger) and a large decrease in the VE (0.87 eV lower). The dihedral angle $\text{O}_{3'}\text{C}_2'\text{N}_1'\text{C}_4'$ changes by 7° , and the angle $\text{C}_2'\text{N}_1'\text{C}_2\text{N}_3$ is practically not affected. In the neutral molecule, the $n_{\text{O}} \rightarrow \pi_{\text{CO}}^*$ excitation characterizes the lowest excited state, without CTIL contributions. This electronic transition is relevant in the chemiluminescence of molecular systems that do not include any good electron-donating part, for instance, 1,2-dioxetane and dioxetanone.^{2,3} Such molecules are much less efficient for chemiluminescence, since the concerted CTIL mechanism does not apply. The second effect studied in the anion, that is, geometry optimizations calculated with the CASPT2 method instead of CASSCF, results in a large stabilization of the CS. The VE is found close to the energy obtained with the big active space (see Table 4). In contrast to the CASSCF geometry optimization, the CASPT2 method provides a structure in which the acetamido part is rotated with respect to the plane of the methylpyrazine group (the dihedral angle $\text{C}_2'\text{N}_1'\text{C}_2\text{N}_3$ is 20°). Much less affected are the CO bond length (0.05 Å larger) and $\text{O}_{3'}\text{C}_2'\text{N}_1'\text{C}_4'$ dihedral angle (5° less negative). Interestingly, the CASPT2 geometry optimization of the pure CTIL state results in a CS with a high efficiency of chemiluminescence ($f = 0.139$). The relative orientation of the acetamido and methylpyrazine parts of the molecule could be the cause for enhancing the emission with respect to the $\text{CS}_{10\text{in}9}$ structure, since the dihedral angle $\text{C}_2'\text{N}_1'\text{C}_2\text{N}_3$ is larger in $\text{CS}_{\text{caspt2}}$ and becomes closer to the values found for the $(S_0)_{\text{min}}$ and $(S_2)_{\text{min}1}$ minima, which show the largest f among all of the structures studied. According to these results, the dihedral angle $\text{C}_2'\text{N}_1'\text{C}_2\text{N}_3$ might be relevant in this molecule to control the interaction between π states. CASPT2(12-in-10) computations at the geometry of the $\text{CS}_{\text{caspt2}}$ result in a lower f (in particular, $f = 0.063$), since the $n_{\text{O}} \rightarrow n_{\text{CO}}^*$ excitation, which corresponds to a symmetry-forbidden electronic transition, is now contributing to the CS. To analyze the relative contribution of the CTIL and $n_{\text{O}} \rightarrow n_{\text{CO}}^*$ excitations in the CS, more accurate CASSCF/CASPT2 studies should be

done with small systems in which the concerted CTIL mechanism applies and all of the relevant orbitals of the dioxetanone moiety (see ref 3) as well as those of the π -aromatic system could be included in the active space.

Apart from the differences found in the geometries obtained by means of the mentioned CASSCF/CASPT2 strategies, all of the optimized structures are characterized by CTIL character, a single bond in the CO group (larger CO distances than 1.28 Å) and sp^3 hybridization of the $\text{C}_{3'}$ atom (lower absolute values than 132° for the dihedral angle $\text{O}_{3'}\text{C}_2'\text{N}_1'\text{C}_4'$). These findings are markedly different from the results obtained in the fluorescence study. While the FS is characterized by a double bond and planar structure for the carbonyl group, the single bond CO and sp^3 hybridization of the $\text{C}_{3'}$ atom are the main features of the CS. Table 4 also includes data reported in the literature related to the chemiluminescence of 2-methanamidopyrazine,¹⁴ which is a small model of coelenteramide close to the molecule studied in the present work. The structures obtained in that study, by optimizing the first excited state of different conformers at the CIS/6-31+G(d) level of theory, were associated with the light-emitting species of this molecule in the chemiluminescence reaction. However, all of the conformers have CO bond lengths around 1.20 Å, and the $\text{O}_{3'}\text{C}_2'\text{N}_1'\text{C}_4'$ dihedral angle is close to 180° (planar conformation of the carbonyl group). These results are in agreement with the FS values obtained for the 2-acetamido-3-methylpyrazine anion and largely differ from the CS patterns established in the present work. Regarding the VE energies, larger values (around 3 eV) are found by Isobe et al. for the lowest-excited state optimized structure,¹⁴ which are closer to the FS than the CS present results. Experimentally, the bioluminescence of aequorin and obelin are measured in the ranges 2.19–2.64 and 2.23–3.09 eV,²³ respectively, and the chemiluminescence of coelenteramide seems to have a value of 2.66 eV.¹⁸ On the other hand, the emission maxima in the chemiluminescence spectra of different imidazopyrazinone analogues measured by Hirano et al. are above 2.52 eV.¹⁵ All of these data are larger than the present VE obtained with the most accurate strategies CASSCF(12-in-10) and CASPT2. As explained in the absorption and fluorescence analysis, the remaining part of the coelenteramide molecule that is not present in 2-acetamido-3-methylpyrazine, mainly the phenolate ring, will contribute to enlarge the energy separation between the CS and ground states. In agreement with the present theoretical findings, the data obtained by Hirano et al. show in all cases larger emission maxima for the fluorescence process with respect to chemiluminescence.¹⁵

Finally, according to the findings of our computations and the associated analysis, some recommendations can be made for theoretical and experimental studies on chemiluminescence. Reliable theoretical studies on the chemiluminescence properties of light-emitting species should not be performed in general by using the ground-state equilibrium structure of the chemiluminescence product as the starting point to find the CS, since different excited state structures may be found for the FS and CS, as shown in the present work. Instead, the whole chemiluminescence reaction must be modeled, from the peroxy precursor to the formation of the emissive state, or the CS can be reached by carrying out geometry optimizations from a structure close to the peroxy intermediate. Regarding the experiments, procedures analyzing the luminescence emitted by the reaction product after irradiation should be avoided in favor of techniques that involve carrying out the whole chemiluminescence reaction and the analysis of the radiation produced. The differences between the

CS and FS might become even stronger in large systems, since the complexity of the PESs increases along with the number of internal degrees of freedom, and therefore the relative positions of the CS and FS states on the excited state surface could differ substantially.

CONCLUSIONS

The characterization of the chemiluminescence and fluorescence states of a small model for the coelenteramide and *Cypridina* oxyluciferin systems, in particular, the 2-acetamido-3-methylpyrazine anion molecule, has been carried out by means of the CASSCF/CASPT2 method in order to establish the molecular basis of the differences between both luminescent phenomena. The neutral form of the molecule has been also considered in some parts of the study, since this protonated form could be also relevant in the chemiluminescence process. In the Franck–Condon region of the ground-state equilibrium structures, the lowest excited states correspond to excitations between orbitals delocalized over the π -conjugated system, while electronic transitions to the carbonile moiety, which are related to the concerted CTIL mechanism responsible for an efficient chemiluminescence process, are found to be very high in energy (above 7 eV, which is more than 2 eV higher in energy with respect to the brightest transitions). In contrast, the CTIL state has been obtained as the lowest-lying excited state of the anion when a geometry close to the peroxy intermediate of the chemiluminescence reaction is used. Therefore, this strategy seems to be more appropriate for finding the CTIL transition and has been employed to characterize the CS. The results obtained for the neutral form by using the same procedure are different: none of the lowest excited states correspond to a CTIL transition. Thus, the anion form is the relevant species for the efficient chemiluminescence of 2-acetamido-3-methylpyrazine, which is consistent with the fact that this deprotonated system has a better activator group (with low ionization potential) for turning on the concerted CTIL mechanism. The protonated species cannot be discarded, however, as the light-emitting molecule in different imidazopyrazinone derivatives in which other chemical groups of the system can act as activator groups, as has been proved elsewhere.^{12,15} In addition, the protonation state of the $N_{1'}$ atom will depend on the closest amino acids of the apoprotein in the bioluminescence process, and the substrate–apoprotein interactions must be analyzed to ultimately understand the role of the protonated/deprotonated species in the chemiluminescence taking place in living organisms. The present photochemical study of the anion molecule has led to three structures which contribute to the fluorescent properties of the molecule: two shallow minima on the S_2 surface and a minimum in the S_1 surface, with VEs of 3.59, 2.21, and 2.11 eV, respectively. The deactivation of the brightest state of the anion molecule is guided by a rotation of the acetamido moiety with respect to the plane of the methylpyrazine part, which reaches a perpendicular conformation at the CI between the S_1 and S_2 surfaces. All of the emissive states found for the anion and neutral systems are characterized by a planar structure (sp^2 hybridization) and double bond of the CO, which do not differ much from the properties of the carbonile group in the ground-state equilibrium minimum. In the analysis of the CS for the anion, distinct multiconfigurational approaches have been employed, finding decreases of the VE when the oxygen lone pair of the CO chemical group is included in the active space and when

dynamical correlation is included in the CASPT2 geometry optimizations. According to these results, the chemiluminescent emission of 2-acetamido-3-methylpyrazine is lower in energy than fluorescence, the most accurate obtained values being in the range 1.60–1.72 eV. All levels of theory result in an out-of-plane conformation of the CO group (sp^3 hybridization) with a large bond length (single bond character), which have been established as the structural patterns of the chemiluminescence phenomenon. The comparison of the results obtained in the fluorescence and chemiluminescence analysis shows significant differences in the electronic structure and geometrical parameters of the CS and FS states for the 2-acetamido-3-methylpyrazine molecule, which must be taken into account in the design of future theoretical and experimental studies on bioluminescence and chemiluminescence. In the light of the present findings, we do not recommend, in general, procedures in which the properties of the chemiluminescent state are obtained by using the ground-state equilibrium structure of the reaction product, unless a previous analysis supporting this strategy had been carried out. A much safer computational approach is to start with a structure close to the peroxy intermediate of the chemical reaction in order to characterize the CS.

ASSOCIATED CONTENT

S Supporting Information. Complete sets of CASSCF orbitals of 2-acetamido-3-methylpyrazine anion and neutral molecules are presented in Figures S1 and S2, respectively. The CASSCF MEPs related to the Figures 4 and 5 and corresponding to the MEP of the S_2 state from the $(S_0)_{\text{min}}$ in the first case, and the MEP of the S_1 state from the CI in the latter, are displayed in Figures S3 and S4, respectively. Additional MEPs on the S_1 surface are also provided: Figures S5–S8 show the CASSCF MEP from the $(S_2)_{\text{min1}}$, the CASPT2//CASSCF MEP from the $(S_2)_{\text{min1}}$, the CASSCF MEP from the $(S_0)_{\text{min}}$ and the CASPT2//CASSCF MEP from the $(S_0)_{\text{min}}$, respectively. The coordinates for the main structures of the present study can be found in Table S1. This material is available free of charge via the Internet at <http://pubs.acs.org>.

AUTHOR INFORMATION

Corresponding Author

*E-mail: Daniel.Roca@kvac.uu.se, Roland.Lindh@kvac.uu.se.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENT

D.R.-S. thanks the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013)/ERC grant agreements no. 255363. I.N. thanks Prof H. M. Marques for funding through the DST/NRF SARChI initiative. Y.-J.L. thanks the National Nature Science Foundation of China. N.F. thanks the French National Center for Scientific Research. R.L. thanks the Swedish Research Council for financial support.

REFERENCES

(1) Nic, M.; Jirat, J.; Kosata, B. *IUPAC. Compendium of Chemical Terminology*, 2nd ed. (the "Gold Book"). <http://goldbook.iupac.org>

(accessed Sep 6, 2011). Compiled by McNaught, A. D.; Wilkinson, A. Blackwell Scientific Publications: Oxford, U. K., 1997. Updates compiled by Jenkins, A. ISBN 0-9678550-9-8. DOI: 10.1351/goldbook.

(2) De Vico, L.; Liu, Y. J.; Krogh, J. W.; Lindh, R. *J. Phys. Chem. A* **2007**, *111*, 8013–8019.

(3) Liu, F.; Liu, Y. J.; De Vico, L.; Lindh, R. *J. Am. Chem. Soc.* **2009**, *131*, 6181–6188.

(4) Liu, F. Y.; Liu, Y. J.; De Vico, L.; Lindh, R. *Chem. Phys. Lett.* **2009**, *484*, 69–75.

(5) Navizet, I.; Liu, Y.-J.; Ferré, N.; Roca-Sanjuán, D.; Lindh, R. *Chem. Phys. Chem.* **2011**, DOI: 10.1002/cphc.201100504.

(6) Seliger, H. H.; McElroy, W. D. *Arch. Biochem. Biophys.* **1960**, *88*, 136–141.

(7) Hawronskyj, J. M.; Holah, J. *Trends Food Sci. Tech.* **1997**, *8*, 79–84.

(8) Lang, T.; Goyard, S.; Lebastard, M.; Milon, G. *Cell. Microbiol.* **2005**, *7*, 383–392.

(9) Gundermann, K.-D. *Chemilumineszenz Organischer Verbindungen*; Springer: Berlin, 1968; pp 1–174.

(10) Gundermann, K.-D. In *Photochemistry. Topics in Current Chemistry*; Boschke, F., Ed.; Springer-Verlag: New York, 1974; Vol. 46; pp 61–139.

(11) *Bioluminescence & Chemiluminescence: Progress and perspectives*, Proceedings of the 13th International Symposium Pacifico Yokohama, Japan, August 2–6; Tsuji, A., Matsumoto, M., Maeda, M., Kricka, L. J., Stanley, P. E., Eds.; World Scientific: Singapore, 2004.

(12) Isobe, H.; Takano, Y.; Okumura, M.; Kuramitsu, S.; Yamaguchi, K. *J. Am. Chem. Soc.* **2005**, *127*, 8667–8679.

(13) Roca-Sanjuán, D.; Aquilante, F.; Lindh, R. *WIREs Comp. Mol. Sci.* **2011**, DOI: 10.1002/wcms.97.

(14) Isobe, H.; Yamanaka, S.; Kuramitsu, S.; Yamaguchi, K. *J. Am. Chem. Soc.* **2008**, *130*, 132–149.

(15) Hirano, T.; Takahashi, Y.; Kondo, H.; Maki, S.; Kojima, S.; Ikeda, H.; Niwa, H. *Photochem. Photobiol. Sci.* **2008**, *7*, 197–207.

(16) Imai, Y.; Shibata, T.; Maki, S.; Niwa, H.; Ohashi, M.; Hirano, T. *J. Photochem. Photobiol. A* **2001**, *146*, 95–107.

(17) Mori, K.; Maki, S.; Niwa, H.; Ikeda, H.; Hirano, T. *Tetrahedron* **2006**, *62*, 6272–6288.

(18) Shimomura, O.; Teranishi, K. *Luminescence* **2000**, *15*, 51–58.

(19) Hirano, T.; Ohmiya, Y.; Maki, S.; Niwa, H.; Ohashi, M. *Tetrahedron Lett.* **1998**, *39*, 5541–5544.

(20) Wu, C.; Nakamura, H.; Murai, A.; Shimomura, O. *Tetrahedron Lett.* **2001**, *42*, 2997–3000.

(21) Fujimori, K.; Komiyama, T.; Tabata, H.; Nojima, T.; Ishiguro, K.; Sawaki, Y.; Tatsuzawa, H.; Nakano, M. *Photochem. Photobiol.* **1998**, *68*, 143–149.

(22) Stepanyuk, G. A.; Golz, S.; Markova, S. V.; Frank, L. A.; Lee, J.; Vysotski, E. S. *FEBS Lett.* **2005**, *579*, 1008–1014.

(23) Belogurova, N. V.; Kudryasheva, N. S.; Alieva, R. R.; Sizykh, A. G. *J. Photochem. Photobiol. B* **2008**, *92*, 117–122.

(24) Shimomura, O.; Johnson, F. H.; Saiga, Y. *J. Cell. Comp. Physiol.* **1962**, *59*, 223–240.

(25) Morin, J. G.; Hastings, J. W. *J. Cell. Physiol.* **1971**, *77*, 305–312.

(26) Shimomura, O.; Johnson, F. H.; Masugi, T. *Science* **1969**, *164*, 1299–1300.

(27) Liu, Z. J.; Stepanyuk, G. A.; Vysotski, E. S.; Lee, J.; Markova, S. V.; Malikova, N. P.; Wang, B. C. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 2570–2575.

(28) Roos, B. O.; Lindh, R.; Malmqvist, P.-Å.; Veryazov, V.; Widmark, P.-O. *J. Phys. Chem. A* **2004**, *108*, 2851–2858.

(29) Andersson, K.; Malmqvist, P.-Å.; Roos, B. O. *J. Chem. Phys.* **1992**, *96*, 1218–1226.

(30) De Vico, L.; Olivucci, M.; Lindh, R. *J. Chem. Theory Comput.* **2005**, *1*, 1029–1037.

(31) Aquilante, F.; De Vico, L.; Ferré, N.; Ghigo, G.; Malmqvist, P.-Å.; Neogrády, P.; Pedersen, T. B.; Pitoňák, M.; Reiher, M.; Roos, B. O.; Serrano-Andrés, L.; Urban, M.; Veryazov, V.; Lindh, R. *J. Comput. Chem.* **2010**, *31*, 224–247.

(32) González-Ramírez, I.; Roca-Sanjuán, D.; Climent, T.; Serrano-Pérez, J. J.; Merchán, M.; Serrano-Andrés, L. *Theor. Chem. Acc.* **2011**, *128*, 705–711.

(33) Bernardi, F.; Olivucci, M.; Robb, M. A. *Pure Appl. Chem.* **1995**, *67*, 783–789.

(34) Roca-Sanjuán, D.; Olaso-González, G.; González-Ramírez, I.; Serrano-Andrés, L.; Merchán, M. *J. Am. Chem. Soc.* **2008**, *130*, 10768–10779.

Pople Style Basis Sets for the Calculation of NMR Spin–Spin Coupling Constants: the 6-31G-J and 6-311G-J Basis Sets

Hanna Kjær* and Stephan P. A. Sauer

Department of Chemistry, University of Copenhagen, DK-2100 Copenhagen Ø, Denmark

Supporting Information

ABSTRACT: We present a modification of the small and popular Pople basis sets, 6-31G and 6-311G, for density functional theory calculations of Fermi contact dominated NMR indirect nuclear spin–spin coupling constants. These new basis sets, 6-31G-J and 6-311G-J, contain twice the number of contracted s-type functions but the same number of contracted p-type functions as the original Pople basis set. For our test set of 12 one-, two- and three-bond coupling constants, the new basis sets augmented with the standard diffuse and polarization functions, i.e., the 6-31+G*-J and 6-311++G**⁻-J basis sets, lead to a maximum deviation of 5 and 2 Hz, respectively, compared to results obtained with the 6 or more times larger aug-pcJ-4 basis set. In correlated wave function calculations using the second-order polarization propagator approximation, the deviations with respect to the aug-ccJ-pVQZ basis set are 8 and 3 Hz and thus slightly larger.

1. INTRODUCTION

It is well-known, that standard one-electron Gaussian basis sets are in general not suited for the calculation of NMR indirect nuclear spin–spin coupling constants (SSCC), while basis sets specially optimized for SSCCs are often too large for routine calculations on larger organic or biological molecules.^{1–18} In the Ramsey formulation,¹⁹ the SSCC is the sum of four different contributions, the Fermi contact (FC), the spin dipolar (SD), the paramagnetic spin–orbit (PSO), and the diamagnetic spin–orbit (DSO) terms. Standard basis sets are typically optimized for energies and are therefore not necessarily optimal for calculations of other properties, which is especially true for SSCCs. Calculations of SSCCs demand basis sets with higher exponents than included in standard basis sets. In particular s-type functions with very high exponents are required for the FC term, as it depends on the electron density at the nucleus. The PSO term requires additional tight p-type functions and the SD term tight p-, d- and f-type functions. The FC term is often the dominant contribution, especially for one-bond couplings between less electronegative atoms, and several basis sets have thus been made to improve the description of the FC term. This includes the cc-pVXZ-Cs and cc-pVXZ-sun basis sets⁵ and the HuzIV-su4^{20–24} and the aug-cc-pVTZ-J^{6,8–10,17,18} basis sets, which all include additional tight s-type functions. In addition there are also basis sets designed to describe all the contributions to the coupling constants accurately. This includes the pcJ-n, aug-pcJ-n^{14,16} and ccJ-pVXZ¹⁵ families of basis sets, which include tight s-, p-, d-, and sometimes even f-type functions. The pcJ-n and aug-pcJ-n basis sets are developed for calculations at the density functional theory (DFT) level, while the ccJ-pVXZ basis sets are designed for correlated wave function methods.

Unfortunately all of these basis sets are rather large; the pcJ-2 basis set, e.g., is of approximately the same size as the aug-pVTZ-J basis set. Consequently these basis sets cannot be used routinely for larger systems like larger organic or biological molecules. Instead, much smaller basis sets, such as the very popular Pople

style^{25,26} or Ahlrichs TZVP and qzp basis sets,^{27,28} are used for larger molecules, even though these basis sets are clearly not adequate for the calculation of SSCCs.

In this study, we will present and test improvements to the Pople basis sets 6-31G and 6-311G for the calculation of FC dominated coupling constants. We only address FC dominated couplings, as most organic and biological couplings are FC dominated. Also, when we only consider FC dominated couplings, we primarily have to focus on the addition of tight s-type functions. Therefore we do not attempt to generate the best possible SSCC basis set, because such basis sets already exist, for example, the aug-pcJ-4 basis set. On the contrary, we want to generate a basis set, which gives acceptable results while still being reasonably small and which can consistently be used together with the Pople basis sets 6-31G and 6-311G. These basis sets should be suitable for organic or biological systems, where larger basis sets cannot be used in routine applications. Finally one should mention that Barone and co-workers²⁹ have recently presented in a similar study new basis sets for the calculation of electron paramagnetic resonance hyperfine coupling constants at the B3LYP level, called N07D. The N07D basis sets are also based on the 6-31G basis sets but are augmented with only one set of core–valence s-, p-, and d-type functions.

We restrict ourselves here to basis sets for hydrogen, carbon, nitrogen, and oxygen to be used in calculations of ¹J(CH), ¹J(NH), ¹J(CN), ¹J(CC), ¹J(CO), ¹J(OH), ²J(HH), and ³J(HH) couplings, as these are the most important FC dominated couplings relevant for organic and biological systems. However, one should note that the ¹J(CO), ¹J(OH), ²J(HH), and ³J(HH) couplings are not completely dominated by the FC term but have also significant contributions from primarily the PSO term as well as non-negligible contributions from the SD and DSO terms. For the development of the basis sets we have studied the

Received: August 4, 2011

Published: October 24, 2011

Table 1. Composition of the Basis Sets Discussed in This Paper

	hydrogen		C, N, O	
	primitives	contracted	primitives	contracted
6-31G	4s	2s	10s4p	3s2p
6-31G-J	7s	4s	13s5p	6s2p
6-31+G*-J	7s	4s	14s6p1d	7s3p1d
6-311G	5s	3s	11s5p	4s3p
6-311G-J	8s	6s	14s6p	8s3p
6-311++G** ⁻ J	9s1p	7s1p	15s7p1d	9s4p1d
pcJ-1*	6s1p	4s1p	9s5p1d	5s4p1d
pcJ-1	6s2p	4s2p	9s5p2d	5s4p2d
pcJ-2*	8s2p1d	5s2p1d	12s7p2d1f	7s5p2d1f
aug-cc-pVTZ-J	13s3p1d	6s3p1d	21s8p3d1f	9s5p3d1f
pcJ-2	8s3p2d	5s3p2d	12s7p3d2f	7s5p3d2f
ccJ-pVTZ	7s3p2d	5s3p2d	12s6p3d1f	7s4p3d1f
aug-ccJ-pVQZ	9s5p4d2f	7s5p4d2f	15s8p5d3f2g	10s6p5d3f2g
aug-pcJ-4	13s8p5d3f2g	9s8p5d3f2g	20s13p8d5f3g2h	12s10p8d5f3g2h

corresponding couplings in methane, ethane, ammonia, water, and formamide (FMA), while the final basis sets were also tested in calculations of the $^1J(\text{NH})$ coupling constant of *N*-methylacetamide.

2. COMPUTATIONAL DETAILS

In all calculations of the SSCCs we employed the Dalton program package.³⁰ In the development of the basis set, the calculations were carried out at the DFT level with the B3LYP exchange–correlation functional,^{31–34} while some of the benchmark calculations were performed at the level of the second-order polarization propagator approximation (SOPPA).^{6,35–40} The B3LYP functional was chosen, as this functional had been used for the development of the (aug-)pcJ-*n* family of basis sets¹⁴ and as B3LYP generally gives good results for SSCCs.^{11,23,24,41–51} All DFT calculations will be benchmarked against the aug-pcJ-4 result, as this basis set is the best possible basis set for DFT calculations and all SOPPA calculations against the aug-ccJ-pVQZ basis set. The geometries of all molecules were optimized by B3LYP/6-311++G**.

3. BASIS SET GENERATION

The 6-31G and 6-311G basis sets are very small (see, e.g., Table 1). There are only s-type functions on hydrogen, and for the second-row atoms, there are only s- and p-type functions. Polarization and diffuse functions can be added using the standard notations. We will base our basis set development on the 6-31G and 6-311G basis sets and add polarization and diffuse functions only, when we have obtained the new 6-31G-J and 6-311G-J basis sets. In Section 3.1 we will thus describe how the new 6-31G-J and 6-311G-J basis sets are generated. We will primarily discuss how the 6-31G-J basis sets were made, as the same procedure was used for the 6-311G-J basis sets.

3.1. Uncontracted Basis Sets. To improve the Pople style basis sets, we first decontracted the basis sets completely. For the 6-31G basis set the differences between the results obtained with the contracted (6-31G) and uncontracted (6-31Guc) basis sets are considerable, differing up to 30 Hz, while the differences

Table 2. Exponents of the s- and p-Type Functions Added to the 6-31G and 6-311G Basis Sets

atom	s ₁	s ₂	s ₃	p ₁
6-31G				
H	186.51684197	2789.59214924	62666.02044953	
C	30750.99553738	469897.3134792	10873731.57976	39.81711939
N	42052.27768090	641872.0643321	14841521.17252	60.55463459
O	55205.88661614	841553.7695902	19428536.09300	81.54048672
6-311G				
H	340.31345127	5170.23631942	118753.0892189	
C	46416.99275025	717812.0184378	16876193.13942	121.32953805
N	63798.90182836	988674.6264273	23421322.38384	175.03570330
O	86856.41933057	1341747.480166	31660948.43265	238.82507403

between the results from the 6-311G and 6-311Guc basis sets are smaller with a maximum difference of 15 Hz. Interestingly, the results obtained with the uncontracted basis sets deviate generally much more from the aug-pcJ-4 results than the ones from the contracted basis sets, which implies that the Pople basis sets in their contracted form benefit from a very fortunate error cancelation as far as SSCCs are concerned.

The uncontracted basis sets were then augmented with s-type functions with large exponents until convergence. The exponents were chosen according to

$$a_{i-1} = \frac{(a_i/a_{i+1})^2}{a_{i+1}/a_{i+2}} a_i \quad (1)$$

where a_i , a_{i+1} , and a_{i+2} are the previously largest, second, and third largest exponents, and a_{i-1} is the new exponent. This implies that the ratio between the exponents are increased, and not kept fixed as in the even-tempered or aug-cc-pVTZ-J basis sets.^{6,8–10,17,18} The exponents are listed in Table 2. Addition of the first tight s-type function is very important, contributing up to 25 Hz, while addition of the second and third s-type function is less important and the fourth s-type function no longer improves the results considerably. Afterward the carbon, nitrogen, and oxygen basis sets were augmented with p-type functions until convergence. It was only necessary to add one tight p-type function to the second-row atoms. The addition of three tight s-type functions and one p-type function defines the 6-31G-Juc basis set. The 6-311Guc was improved in the same way leading to the 6-311G-Juc basis set, however, is it worth noticing that the two basis sets, 6-31G-Juc and 6-311G-Juc, give almost the same results, the largest difference is 1 Hz. In this context it is important to stress that we did not only consider the total SSCC but also the four individual contributions and that we found that the good correspondence between the results obtained with the uncontracted 6-31G-Juc and 6-311G-Juc basis sets is not due to fortunate error cancelations but to a good description of all four terms.

3.2. Contracted Basis Sets. The optimal uncontracted basis sets were then contracted again. Contrary to the aug-cc-pVTZ-J basis sets,^{6,8–10,17,18} we did not employ molecular orbital coefficients of the simplest hydrides of each atom but natural orbital coefficients from complete active space (CAS) calculations on the atoms. For the second-row atoms all electrons were included in the CAS-space, while all orbitals were included for hydrogen. However, the orbital coefficients of the relevant orbitals obtained from the CAS calculations do not differ significantly from those

Table 3. DFT/B3LYP Results for the Coupling Constants in Hz in the Test Set as a Function of the One-Electron Basis Set Employed

basis	$^1J(\text{CC})$	$^1J(\text{CN})$	$^1J(\text{CO})$	$^1J(\text{CH})$		$^1J(\text{NH})$		$^1J(\text{OH})$	$^2J(\text{HH})$		$^3J(\text{HH})$	
	C_2H_6	FMA	FMA	CH_4	FMA	NH_3	FMA	H_2O	CH_4	FMA	FMA	C_2H_6
aug-pcJ-4	34.62	-14.03	28.96	134.32	201.68	-66.82	-97.72	-76.87	-13.33	7.51	15.57	18.25
pcJ-2	35.18	-13.86	28.91	135.07	199.53	-65.78	-96.71	-74.63	-13.73	7.47	15.42	18.01
aug-cc-pVTZ-J	34.58	-14.16	28.93	134.65	202.35	-66.75	-97.84	-76.15	-13.54	7.30	15.45	18.13
pcJ-2*	35.09	-13.86	29.11	134.99	199.44	-65.71	-96.64	-74.46	-13.79	7.36	15.39	17.98
pcJ-1	37.00	-15.35	28.10	133.88	196.67	-63.89	-96.19	-69.86	-13.50	7.02	14.67	17.58
pcJ-1*	36.92	-15.35	28.31	133.85	196.65	-63.87	-96.17	-69.75	-13.56	6.89	14.64	17.57
6-31G	46.71	-15.32	21.97	140.52	196.09	-62.36	-88.97	-53.66	-18.12	-0.64	11.84	14.77
6-31Guc	31.34	-11.58	39.13	111.34	165.71	-51.79	-80.70	-43.58	-12.13	3.33	10.71	13.14
6-31+G*	48.45	-18.99	7.46	138.45	197.31	-63.32	-88.89	-70.89	-15.19	2.34	11.50	14.01
6-31++G**	48.34	-19.61	7.52	138.98	200.20	-66.91	-92.77	-83.17	-14.27	2.69	11.34	14.08
6-31G-J	36.75	-14.15	44.35	137.19	204.39	-64.28	-99.81	-51.67	-16.72	3.16	13.40	16.75
6-31G-Juc	35.93	-13.49	43.67	136.12	202.66	-63.44	-98.70	-53.24	-15.60	4.69	14.52	17.60
6-31+G*-J	35.98	-15.18	29.34	136.05	204.43	-66.79	-99.13	-72.14	-15.72	4.62	13.91	16.79
6-31++G** ⁻ J	35.47	-14.92	29.43	134.97	204.10	-66.76	-98.69	-73.71	-14.73	5.19	14.02	17.08
6-311G	37.75	-14.40	42.88	129.15	187.45	-60.45	-92.49	-52.70	-15.30	0.81	11.25	14.60
6-311Guc	32.35	-11.86	39.55	117.36	174.50	-55.62	-85.40	-47.99	-13.06	3.55	11.29	14.21
6-311+G*	37.60	-15.35	26.55	125.86	181.64	-60.65	-89.32	-64.56	-13.44	3.71	11.91	14.55
6-311++G**	34.73	-14.46	27.42	119.85	180.07	-59.29	-87.84	-66.72	-11.84	4.55	11.88	14.53
6-311G-J	36.19	-13.82	42.15	136.96	203.08	-64.79	-99.37	-54.39	-15.81	4.30	13.59	17.05
6-311G-Juc	35.70	-13.39	42.69	136.16	202.71	-64.28	-98.80	-54.12	-15.58	4.83	14.14	17.53
6-311+G*-J	35.37	-14.60	29.38	135.81	203.06	-66.27	-98.33	-69.17	-15.08	5.48	14.70	17.54
6-311++G** ⁻ J	35.00	-14.26	29.48	135.51	203.42	-67.02	-98.60	-75.07	-14.05	6.11	14.78	17.63

obtained with Hartree–Fock calculations on the molecules. Nevertheless, we chose the atomic orbital coefficients as these coefficients are uniquely defined.

First we tried a contraction scheme for the 6-31G-Juc basis set, which would be equivalent to the original 6-31G contraction. However, contracting the s-type functions as in the 6-31G basis set gave unacceptably large errors. Jensen had also observed that the s-space cannot be contracted much, without loss of accuracy.¹⁴ Therefore we used the contraction scheme shown in Table 1. Unfortunately, the same contraction scheme could not be applied to the 6-311G-Juc basis set, and therefore the 6-311G-J basis sets are larger relative to the 6-311G basis sets than 6-31G-J is to 6-31G, as seen in Table 1. In both cases the new basis sets have twice the number of contracted s-type function as the original Pople basis sets but the same number of contracted p-type functions. The contracted basis sets can be found in the Supporting Information. The error introduced by the contraction is up to 2 Hz for 6-31G-J and only up to 1 Hz for 6-311G-J. Consequently, due to the contraction, the 6-311G-J basis set performs slightly better than the 6-31G-J basis set.

4. BENCHMARKING THE NEW BASIS SETS

4.1. DFT Calculations. The new basis sets were employed in calculations of the SSCCs of water, $^1J(\text{OH})$, methane, $^1J(\text{CH})$ and $^2J(\text{HH})$, ammonia, $^1J(\text{NH})$, ethane, $^1J(\text{CC})$ and $^3J(\text{HH})$, and formamide, (CH_3NO , FMA), $^1J(\text{CH})$, $^1J(\text{NH})$, $^1J(\text{CN})$, $^1J(\text{CO})$, $^2J(\text{HH})$, and $^3J(\text{HH})$; in total 12 one-, two-, and three-bond coupling constants. We compare the results obtained with the pcJ-2, aug-cc-pVTZ-J, pcJ-2*, pcJ-1, pcJ-1*, 6-31G, 6-31Guc,

Table 4. Mean (Absolute) Errors in Hz and % of the DFT/B3LYP Calculated Coupling Constants in the Test Set Relative to the Corresponding Results Obtained with the aug-pcJ-4 Basis Set

basis	size ^a	CH_3NO	ME (Hz)	MAE (Hz)	ME (%)	MAE (%)
pcJ-2	225		-0.23	0.74	0.46	1.34
aug-cc-pVTZ-J	198		-0.07	0.23	0.19	0.73
pcJ-2*	165		-0.23	0.78	0.58	1.56
pcJ-1	111		-0.33	1.98	1.60	4.54
pcJ-1*	87		-0.33	1.99	1.71	4.69
6-31G	33		-1.72	7.40	11.63	25.75
6-31Guc	78		-0.23	12.96	19.49	25.34
6-31+G*	60		0.83	6.87	10.36	25.75
6-31++G**	72		2.14	6.03	7.98	24.36
6-31G-J	48		-3.10	5.37	2.12	17.33
6-31G-Juc	105		-3.21	4.51	1.09	13.50
6-31+G*-J	75		0.00	1.83	2.32	8.25
6-31++G** ⁻ J	87		0.06	1.33	2.32	6.26
6-311G	48		-1.37	7.44	9.97	22.40
6-311Guc	93		-0.56	10.34	15.24	21.33
6-311+G*	75		1.14	6.11	11.16	14.31
6-311++G**	87		1.61	6.46	12.72	13.28
6-311G-J	69		-2.75	4.50	1.90	14.08
6-311G-Juc	120		-2.95	4.32	1.37	13.11
6-311+G*-J	96		-0.48	1.57	2.00	5.88
6-311++G** ⁻ J	108		-0.07	0.87	1.38	3.54

^aTotal number of basis functions for formamide in the given basis set. The aug-pcJ-4 basis set contains 789 basis functions for formamide.

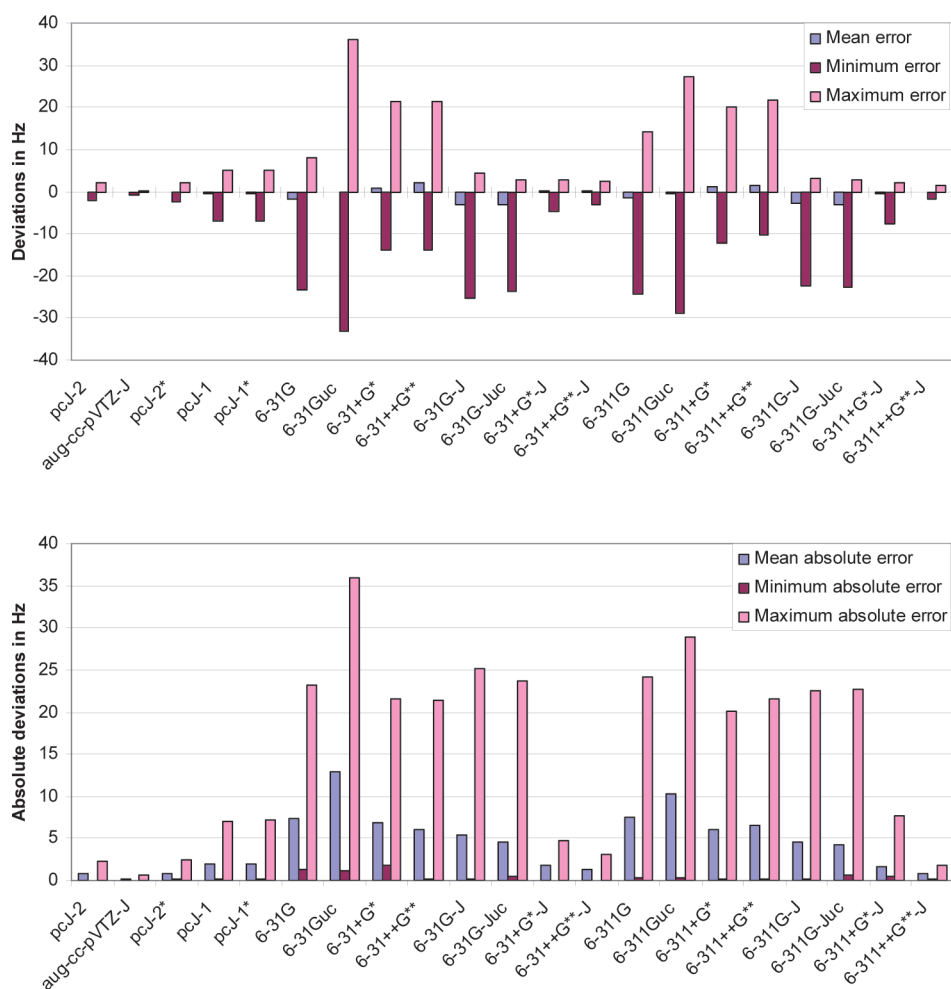


Figure 1. Mean, minimum, and maximum (top) and mean absolute, minimum absolute, and maximum absolute (bottom) errors in Hz of the SSCCs calculated with the given basis set relative to the aug-pcJ-4 result.

6-31+G*, 6-31++G**, 6-31G-J, 6-31G-Juc, 6-31+G*-J, 6-31++G**-J, 6-311G, 6-311Guc, 6-311+G*, 6-311++G**, 6-311G-J, 6-311G-Juc, 6-311+G*-J, and 6-311++G**-J basis sets with the corresponding results obtained with the aug-pcJ-4 basis set. We have chosen to use the aug-pcJ-4 results as reference, as this is the best possible basis set for DFT calculations of SSCCs that we are aware of. We include pcJ-2 in the comparison, as this is recommended for practical calculations,¹⁴ and aug-cc-pVTZ-J, as it is of approximately the same size as pcJ-2, but optimized for FC dominated couplings. We also include a pruned pcJ-2 basis set, pcJ-2*, where the tight p- and d-type functions are removed on hydrogen and the tight d- and f-type functions on non-hydrogen atoms, as these tight functions are not expected to be important for FC dominated couplings. We also include the pcJ-1 basis set, which is of approximately double- ζ quality, as well as its pruned counterpart pcJ-1*, which is of approximately the same size as our new Pople style basis sets.

Finally we include the new basis sets and their corresponding original Pople basis sets. Diffuse functions as well as polarization functions are expected to be important, and we added these functions to our basis sets, using the standard notation. The results are shown in Table 3, and the mean, maximum, and minimum deviations for each basis set are shown in Table 4 and in Figures 1 and 2.

The pcJ-2 and aug-cc-pVTZ-J basis sets are of approximately the same size, and both give excellent results for these couplings compared to the aug-pcJ-4 results. Nevertheless, the aug-cc-pVTZ-J basis set is slightly better for these couplings, which could be expected, as it is optimized for FC dominated couplings. For both pcJ-2* and pcJ-1* the results compare excellently with the results of their larger pcJ-2 and pcJ-1 counterparts, as expected for these couplings. The pcJ-1* performs slightly worse than the 6-311++G**-J basis set, which is also slightly larger. Compared to 6-31+G*-J, which contains fewer basis functions, the absolute deviations of the pcJ-1* results are slightly larger, while the relative deviations are smaller. The reason for this apparent discrepancy is that the relative deviations are dominated by the small ${}^2J(\text{HH})$ coupling constants, which are not as well described by the new Pople style basis sets.

For 6-31G we observe large errors, which are not reduced by uncontraction or addition of polarization and diffuse functions. The same trends are seen for 6-311G. For both 6-31G-J and 6-311G-J we see that the basis sets can systematically be improved by addition of polarization and diffuse functions. Especially diffuse and polarization functions on the second-row atoms are mandatory, while the polarization and diffuse functions are less important for hydrogen; the exception being the ${}^1J(\text{OH})$ coupling in combination with the 6-311++G**-J basis set, where

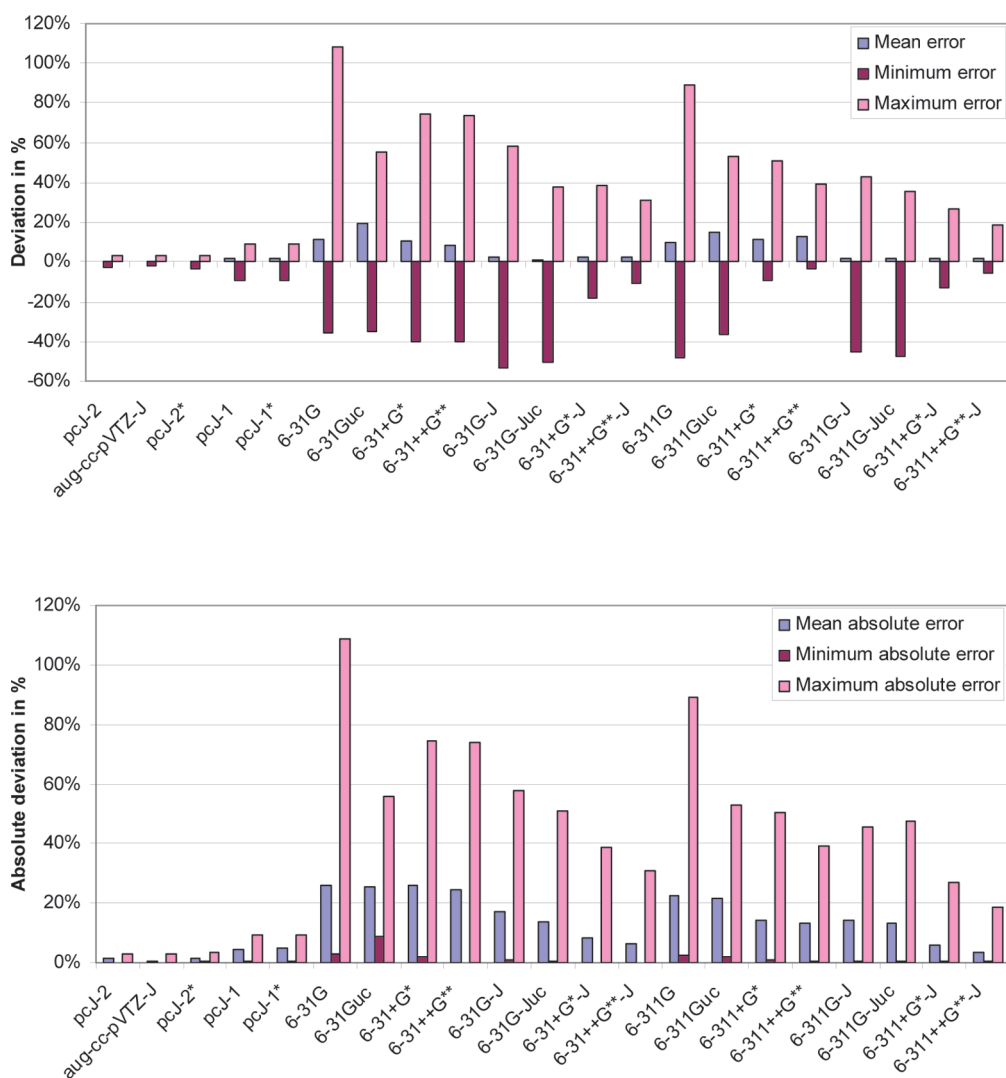


Figure 2. Mean, minimum, and maximum relative errors (top) and mean absolute, minimum absolute, and maximum absolute relative errors (bottom) in % of the SSCCs calculated with the given basis set relative to the aug-pcJ-4 result.

we see a 6 Hz improvement relative to the 6-311+G*-J result. The 6-31+G*-J basis set gives results which deviate less than 5 Hz for all couplings, and 6-311++G***-J gives results which deviate less than 2 Hz for all couplings from the corresponding aug-pcJ-4 results. The largest error is observed for $^1J(\text{OH})$, even though $^1J(\text{CH})$ in formamide exhibits also a large error, however, this coupling is approximately 200 Hz. The relative errors are dominated by $^2J(\text{HH})$ in formamide, which is calculated to be 7.51 Hz by aug-pcJ-4. Here the errors with the new basis sets are between 1.4 and 4.4 Hz leading to the enormous relative errors in the figures. The $^2J(\text{HH})$, the $^3J(\text{HH})$, and the couplings involving oxygen are not completely dominated by the FC term. These couplings are thus not as well described by the new basis sets, even though the largest error produced by the 6-311++G***-J basis set is less than 2 Hz. Consequently these basis sets work well for FC dominated couplings but should be used with caution and only after careful benchmarking against better basis sets for couplings that have important contributions from the PSO, SD, and DSO terms. We have therefore not generated a corresponding basis set for fluorine, as fluorine couplings often

have very large contributions from especially the PSO term.^{10,12,14,44,52-54}

Furthermore we have only investigated the performance of the 6-31+G*-J basis set and not that of the 6-31+G-J and 6-31G*-J basis sets. For some couplings only the diffuse functions are important, as in $^1J(\text{NH})$ in ammonia, while for others the polarization functions are important, as in $^1J(\text{CO})$ in formamide, and for others we need both contributions, as in $^1J(\text{OH})$ in water. Consequently one should always first test the effect of the different diffuse and polarization functions in order to see whether one could employ a smaller basis set.

Finally, we tested the performance of the new 6-31+G*-J basis set also for the $^1J(\text{NH})$ coupling constant of *N*-methylacetamide but did not include it in the test set, although it is also FC dominated. We find that the 6-31+G*-J result deviates by only 1.1 Hz from the corresponding aug-pcJ-4 result.

4.2. Correlated Wave Function Calculations. So far we have only investigated the performance of the new basis sets in DFT calculations and not in calculations with correlated wave function methods. However, the basis set dependence of correlated wave function methods differs from that of DFT. We have therefore

Table 5. Mean (Absolute) Errors in Hz and % of the SOPPA Calculated Coupling Constants in the Test Set Relative to the Corresponding Results Obtained with the aug-ccJ-pVQZ Basis Set

basis	ME (Hz)	MAE (Hz)	ME (%)	MAE (%)
ccJ-pVTZ	-0.12	0.81	0.43	1.67
aug-cc-pVTZ-J	-0.04	0.37	-0.35	1.37
6-31G	-1.08	8.03	13.25	41.44
6-31Guc	-0.24	8.94	15.55	24.59
6-31+G*	1.09	6.73	12.73	31.84
6-31++G**	2.41	6.76	10.56	29.88
6-31G-J	-2.40	7.24	0.39	29.79
6-31G-Juc	-2.64	6.34	-3.32	23.53
6-31+G*-J	0.06	3.12	2.82	14.59
6-31++G**-J	0.42	1.93	3.02	11.06
6-311G	-0.87	4.99	8.22	31.56
6-311Guc	-0.53	6.70	11.53	14.91
6-311+G*	1.10	4.36	10.51	18.62
6-311++G**	1.64	5.06	11.53	14.91
6-311G-J	-2.30	6.43	-1.60	24.63
6-311G-Juc	-2.48	6.22	-3.25	22.61
6-311+G*-J	-0.53	2.56	1.77	10.95
6-311++G**-J	0.01	1.38	1.03	6.70

also carried out the calculations of the previous section at the SOPPA level. The results are given in the Supporting Information. For these calculations we compared to the aug-ccJ-pVQZ basis set¹⁵ because the ccJ-pVXZ basis sets were optimized for correlated wave function calculations in contrast to the pcJ-*n* family of basis sets. The augmentation functions are taken from aug-cc-pVQZ. Due to the computational cost of the ccJ-pVSZ and aug-ccJ-pVSZ, we could not obtain the results of these basis sets for all the investigated molecules. For water and ammonia we could obtain both the aug-ccJ-pVSZ and the ccJ-pVSZ results, and in both cases the aug-ccJ-pVQZ results differed less than 0.1 Hz from the aug-ccJ-pVSZ results, and the aug-ccJ-pVQZ results were closer to the aug-ccJ-pVSZ results than the ccJ-pVSZ results. For methane and ethane where we could obtain the ccJ-pVSZ results, the aug-ccJ-pVQZ differed less than 0.3 Hz from the ccJ-pVSZ results. So it seems that the aug-ccJ-pVQZ results will be converged with respect to the aug-ccJ-pVSZ basis set, and therefore we will employ this in the benchmarking of the new Pople style basis sets. The statistics are shown in Table 5. We observe in general the same trends as for DFT, however, the errors relative to the reference calculations are slightly larger than in the DFT calculations. The 6-31+G*-J results deviate now less than 8 Hz from the aug-ccJ-pVQZ results, while 6-311++G**-J gives results which deviate less than 3 Hz from the aug-ccJ-pVQZ results. They are still superior to the original Pople basis sets, but they are not performing quite as well for SOPPA as they were in the DFT calculations, however, they will still be useful for SOPPA-calculations on large molecules.

5. SUMMARY

For the calculation of NMR indirect nuclear spin-spin coupling constants we present modified 6-31G and 6-311G Pople basis sets. The new basis sets, called 6-31G-J and 6-311G-J, have the same number of contracted p-type functions

and twice the number of contracted s-type functions as the original Pople basis sets. These basis sets can further be augmented with the polarization and diffuse functions of the standard Pople basis sets.

The performance of the new basis sets has been evaluated with DFT/B3LYP and SOPPA calculations on a test set of 12 one-, two- and three-bond coupling constants. We find that polarization and diffuse functions on the second-row atoms are mandatory, while the polarization and diffuse functions on hydrogen are less important, except in the case of ¹J(OH). The 6-31+G*-J basis set gives DFT results that deviate less than 5 Hz from the aug-ccJ-4 results for the 12 couplings investigated in this study, and 6-311++G**-J deviates less than 2 Hz from the aug-ccJ-4 results. At the SOPPA level, the maximum deviations from aug-ccJ-pVQZ results are 8 and 3 Hz, respectively.

These basis sets are only optimized for the calculation of FC dominated couplings and are not expected to give accurate results for couplings where the PSO, DSO or SD contributions will be important. Furthermore we have observed that the couplings involving oxygen and the ²J(HH) couplings are not as accurately described as the other couplings. Therefore these new basis sets are excellent for DFT calculations of FC dominated SSCCs, but they can also be employed in calculations with correlated wave functions. However, in the latter case the inclusion of polarization and diffuse functions on hydrogen is mandatory.

■ ASSOCIATED CONTENT

S Supporting Information. Final contracted 6-31G-J and 6-311G-J basis sets. In addition all SOPPA results for the SSCCs are given. This information is available free of charge via the Internet at <http://pubs.acs.org/>.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: hanna.kjaer@gmail.com.

■ ACKNOWLEDGMENT

The authors thank the Danish Center for Scientific Computing (DCSC) and the Danish Natural Science Research Council/The Danish Councils for Independent Research (grant number 272-08-0486) for financial support.

■ REFERENCES

- (1) Schulman, J. M.; Kaufman, D. N. *J. Chem. Phys.* **1970**, *53*, 477–484.
- (2) Kowalewski, J.; Laaksonen, A.; Roos, B.; Siegbahn, P. *J. Chem. Phys.* **1979**, *71*, 2896–2902.
- (3) Oddershede, J.; Geertsen, J.; Scuseria, G. E. *J. Phys. Chem.* **1988**, *92*, 3056–3059.
- (4) Geertsen, J.; Oddershede, J.; Raynes, W. T.; Scuseria, G. E. *J. Magn. Reson.* **1991**, *93*, 458–471.
- (5) Helgaker, T.; Jaszunski, M.; Ruud, K.; Górska, A. *Theor. Chem. Acc.* **1998**, *99*, 175–182.
- (6) Enevoldsen, T.; Oddershede, J.; Sauer, S. P. A. *Theor. Chem. Acc.* **1998**, *100*, 275–284.
- (7) Guilleme, J.; San Fabián, J. *J. Chem. Phys.* **1998**, *109*, 8168–8181.
- (8) Sauer, S. P. A.; Raynes, W. T. *J. Chem. Phys.* **2000**, *113*, 3121–3129.

- (9) Provasi, P. F.; Aucar, G. A.; Sauer, S. P. A. *J. Chem. Phys.* **2001**, *115*, 1324–1334.
- (10) Barone, V.; Provasi, P. F.; Peralta, J. E.; Snyder, J. P.; Sauer, S. P. A.; Contreras, R. H. *J. Phys. Chem. A* **2003**, *107*, 4748–4754.
- (11) Peralta, J. E.; Scuseria, G. E.; Cheeseman, J. R.; Frisch, M. J. *Chem. Phys. Lett.* **2003**, *375*, 452–458.
- (12) Sanchez, M.; Provasi, P. F.; Aucar, G. A.; Sauer, S. P. A. *Adv. Quantum Chem.* **2005**, *48*, 161–183.
- (13) Deng, W.; Cheeseman, J. R.; Frisch, M. J. *J. Chem. Theory Comput.* **2006**, *2*, 1028–1037.
- (14) Jensen, F. *J. Chem. Theory Comput.* **2006**, *2*, 1360–1369.
- (15) Benedikt, U.; Auer, A. A.; Jensen, F. *J. Chem. Phys.* **2008**, *129*, 064111.
- (16) Jensen, F. *Theor. Chem. Acc.* **2010**, *126*, 371–382.
- (17) Rusakov, Y. Y.; Krivdin, L. B.; Sauer, S. P. A.; Levanova, E. P.; Levkovskaya, G. G. *Magn. Reson. Chem.* **2010**, *48*, 633–637.
- (18) Provasi, P. F.; Sauer, S. P. A. *J. Chem. Phys.* **2010**, *133*, 54308.
- (19) Ramsey, N. F. *Phys. Rev.* **1953**, *91*, 303–307.
- (20) Huzinaga, S. *J. Chem. Phys.* **1965**, *42*, 1293–1302.
- (21) Schindler, M.; Kutzelnigg, W. *J. Chem. Phys.* **1982**, *76*, 1919–1933.
- (22) van Wüllen, C. *Die Berechnung magnetischer Eigenschaften unter Berücksichtigung der Elektronenkorrelation: Die Multikonfigurations-Verallgemeinerung der IGLO-Methode*. Dissertation; Ruhr-Universität Bochum: Bochum, Germany, 1992; pp 54–55.
- (23) Helgaker, T.; Watson, M.; Handy, N. C. *J. Chem. Phys.* **2000**, *113*, 9402–9409.
- (24) Ruden, T. A.; Lutnæs, O. B.; Helgaker, T. *J. Chem. Phys.* **2003**, *118*, 9572–9582.
- (25) Krishnan, R.; Binkley, J. S.; Seeger, R.; Pople, J. A. *J. Chem. Phys.* **1980**, *72*, 650–654.
- (26) Clark, T.; Chandrasekhar, J.; Spitznagel, G. W.; Schleyer, P. V. R. *J. Comput. Chem.* **1983**, *4*, 294–301.
- (27) Schäfer, A.; Horn, H.; Ahlrichs, R. *J. Chem. Phys.* **1992**, *97*, 2571–2577.
- (28) Schäfer, A.; Huber, C.; Ahlrichs, R. *J. Chem. Phys.* **1994**, *100*, 5829–5835.
- (29) Barone, V.; Cimino, P.; Stendardo, E. *J. Chem. Theory Comput.* **2008**, *4*, 751–764.
- (30) *Dalton, a molecular electronic structure program*, release 2.0; 2005; <http://daltonprogram.org/> (accessed October 24, 2011).
- (31) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785–789.
- (32) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648–5652.
- (33) Vosko, S. H.; Wilk, L.; Nusair, M. *Can. J. Phys.* **1980**, *58*, 1200–1211.
- (34) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. *J. Chem. Phys.* **1994**, *98*, 1623–11627.
- (35) Nielsen, E. S.; Jørgensen, P.; Oddershede, J. *J. Chem. Phys.* **1980**, *73*, 6238–6246.
- (36) Nielsen, E. S.; Jørgensen, P.; Oddershede, J. *J. Chem. Phys.* **1981**, *75*, 499.
- (37) Geertsen, J.; Oddershede, J. *Chem. Phys.* **1984**, *90*, 301–311.
- (38) Sauer, S. P. A.; Diercks, G. H. F.; Oddershede, J. *Int. J. Quantum Chem.* **1991**, *39*, 667–679.
- (39) Packer, M. J.; Dalskov, E. K.; Enevoldsen, T.; Jensen, H. J. A.; Oddershede, J. *J. Chem. Phys.* **1996**, *105*, 5886–5900.
- (40) Bak, K. L.; Koch, H.; Oddershede, J.; Christiansen, O.; Sauer, S. P. A. *J. Chem. Phys.* **2000**, *112*, 4173–4185.
- (41) Bouř, P.; Buděšínský, M. *J. Chem. Phys.* **1999**, *110*, 2836–2843.
- (42) Sychrovský, V.; Gräfenstein, J.; Cremer, D. *J. Chem. Phys.* **2000**, *113*, 3530–3547.
- (43) Peralta, J. E.; Contreras, R. H.; Snyder, J. P. *Chem. Commun.* **2000**, 2025–2026.
- (44) Barone, V.; Peralta, J. E.; Contreras, R. H.; Snyder, J. P. *J. Phys. Chem. A* **2002**, *106*, 5607–5612.
- (45) Lutnæs, O. B.; Ruden, T. A.; Helgaker, T. *Magn. Reson. Chem.* **2004**, *42*, S117–S127.
- (46) Keal, T. W.; Tozer, D. J.; Helgaker, T. *Chem. Phys. Lett.* **2004**, *391*, 374–379.
- (47) Maximoff, S. N.; Peralta, J. E.; Barone, V.; Scuseria, G. E. *J. Chem. Theory Comput.* **2005**, *1*, 541–545.
- (48) Keal, T. W.; Helgaker, T.; Salek, P.; Tozer, D. J. *Chem. Phys. Lett.* **2006**, *425*, 163–166.
- (49) Sauer, S. P. A.; Provasi, P. F. *ChemPhysChem* **2008**, *9*, 1259–1261.
- (50) Provasi, P. F.; Sauer, S. P. A. *Phys. Chem. Chem. Phys.* **2009**, *11*, 3987–3995.
- (51) Møgelhøj, A.; Aidas, K.; Mikkelsen, K. V.; Sauer, S. P. A.; Kongsted, J. *J. Chem. Phys.* **2009**, *130*, 134508.
- (52) Peralta, J. E.; Barone, V.; Contreras, R. H.; Zaccari, D. G.; Snyder, J. P. *J. Am. Chem. Soc.* **2001**, *123*, 9162–9163.
- (53) Provasi, P. F.; Aucar, G. A.; Sauer, S. P. A. *J. Chem. Phys. A* **2004**, *108*, 5393–5398.
- (54) Provasi, P. F.; Sauer, S. P. A. *J. Chem. Theory Comput.* **2006**, *2*, 1019–1027.

Optimized Basis Sets for Calculation of Electron Paramagnetic Resonance Hyperfine Coupling Constants: aug-cc-pVTZ-J for the 3d Atoms Sc–Zn

Erik Donovan Hedegård,[†] Jacob Kongsted,[†] and Stephan P. A. Sauer^{*,‡}

[†]Department of Chemistry and Physics, University of Southern Denmark, Odense, Denmark

[‡]Department of Chemistry, University of Copenhagen, Copenhagen, Denmark

S Supporting Information

ABSTRACT: The hyperfine coupling tensor of electron paramagnetic resonance (EPR), describing the interaction between an electron and a given nuclei, depends strongly on the electron density at the nucleus. With standard Gaussian-type orbital basis sets (GTOs), employed in most calculations, it is difficult to obtain converged results of the hyperfine coupling tensor, and basis sets with more flexible core regions have therefore been devised. To this class of core property basis sets belong the aug-cc-pVTZ-J basis sets developed for the s- and p-block atoms. Here, we extend the aug-cc-pVTZ-J basis sets to include the 3d elements Sc–Zn. The converged optimal basis sets are throughout the series described by a (25s17p10d3f2g)/[17s10p7d3f2g] contraction scheme, where four tight s-, one tight p-, and one tight d-type function have been added to the original aug-cc-pVTZ basis sets. The basis sets are generally contracted, and molecular orbital coefficients are used as contraction coefficients. By validation studies with different functionals and compounds, it is shown that the values of the contraction coefficient are effectively independent of the compound used in their generation and the exchange-correlation functional employed in the calculation.

INTRODUCTION

Electron paramagnetic resonance (EPR) has, since its early development in the 1940s,¹ evolved to be a spectroscopic method of fundamental importance with applications in a wide range of chemical fields, including d-block metals,² organic radicals,^{3–5} and biomimetic chemistry.^{6–8} The experimental spectra are usually interpreted by the introduction of an effective spin Hamiltonian where the physical quantities enter as fitting parameters. Within d-block metal chemistry, the effective spin Hamiltonian was developed with impressing insight by Abragam and Pryce, using a framework of crystal field theory^{9–13}

$$\hat{\mathcal{H}}_{\text{eff}} = \mu_B \hat{\mathbf{S}} \mathbf{g} \mathbf{B} + \sum_{\alpha} \hat{\mathbf{S}} \mathbf{A}_{\alpha} \hat{\mathbf{I}}_{\alpha} + D \left[\hat{S}_z^2 - \frac{1}{3} S(S+1) \right] + E [\hat{S}_x^2 - \hat{S}_y^2] \quad (1)$$

Here, $\hat{\mathbf{S}}$ is the total spin of the molecule and \mathbf{B} is the external magnetic field. The \mathbf{g} tensor is a 3×3 tensor, describing the interaction of the external field with the spin of the unpaired electron(s) in the molecule. D and E enter for $S > 1/2$ systems and are zero field splitting constants which depend on the exact symmetry of the paramagnetic center. The hyperfine coupling tensor, \mathbf{A} , is the target of the present study. As the \mathbf{g} tensor, the hyperfine coupling tensor is a 3×3 tensor. It describes the interaction between the unpaired electron(s) in the molecular system and the magnetic moments of the nuclei, $\hat{\mathbf{I}}_{\alpha}$. Accordingly, the sum in eq 1 runs over the nuclei in the molecule which have nonzero spin. The hyperfine coupling tensor is used as a sensitive probe for the immediate coordination environment of d-block metals, for instance in the studies of the cytochrome P450 enzymes^{14–17} or the active species in photosystem II.¹⁸

From textbooks on the magnetic theory of quantum mechanics^{1,19} it is known that \mathbf{A} , to first order, is composed of an isotropic (Fermi-contact) term and a contribution due to anisotropic spin-dipolar interactions

$$\mathbf{A}_{\text{tot}} = \mathbf{A}_{\text{iso}} + \mathbf{A}_{\text{SD}} \quad (2)$$

The components of the isotropic, \mathbf{A}_{iso} , and anisotropic, \mathbf{A}_{SD} , spin dipolar hyperfine coupling tensors can be calculated from first principles from the N -electron, molecular Born–Oppenheimer wave function, Ψ , as

$$A_{\text{iso}, \alpha\beta} = \frac{\hbar \mu_0}{3m_e} \frac{eg_e g_K \mu_N}{\langle \Psi | \hat{S}_{\alpha} | \Psi \rangle} \langle \Psi | \sum_i \delta(\mathbf{r}_i - \mathbf{R}_K) \hat{s}_{i,\beta} | \Psi \rangle \quad (3)$$

$$A_{\text{SD}, \alpha\beta} = \frac{\hbar \mu_0}{8\pi m_e} \frac{eg_e g_K \mu_N}{\langle \Psi | \hat{S}_{\alpha} | \Psi \rangle} \langle \Psi | \sum_i \frac{3(\hat{s}_i(\mathbf{r}_i - \mathbf{R}_K))_{\beta}}{|\mathbf{r}_i - \mathbf{R}_K|^5} - \frac{\hat{s}_{i,\beta}}{|\mathbf{r}_i - \mathbf{R}_K|^3} | \Psi \rangle \quad (4)$$

where α and β are Cartesian components and the physical constants have their usual meaning.¹⁹ \mathbf{r}_i and \mathbf{R}_K are coordinates for the i th electron and the K th nucleus, respectively, and $\delta(\mathbf{r}_i - \mathbf{R}_K)$ denotes the Dirac δ function in which the expectation value over Ψ returns the electron density at nucleus K . Here, the Dirac δ function is multiplied with the spin operator for the i th electron, $\hat{s}_{i,\beta}$, thus returning the spin density at the K th nucleus. A representation which leads to a diagonal total tensor, \mathbf{A}_{tot} is usually chosen, and this practice is followed in this work. However, *a priori* the

Received: August 21, 2011

Published: November 14, 2011

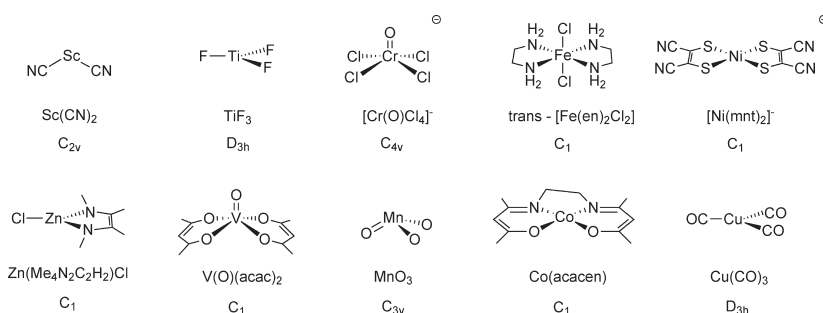


Figure 1. Training set 1 used for construction of the aug-cc-pVTZ-Juc basis set.

tensors are not represented by diagonal matrices. Further contributions arise if spin–orbit coupling is considered,²⁰ but such an effect will be left out in this study.

Calculation of A_{tot} from first principles has unfortunately proven to be a difficult task, and early attempts were met with little success.¹³ Early studies by Watson and Freeman²¹ concluded that the core level spin-polarization was crucial for the correct description of hyperfine couplings. A wide range of studies on small systems in recent decades^{22–33} have shown that highly correlated methods and basis sets with a flexible core region must be invoked to recover this core-level spin polarization. The special basis set requirements arise due to the Dirac δ function in the expression for A_{iso} , which usually is much larger than that for A_{SD} . Accordingly, the basis set should provide an accurate description of the core (and semicore) spin polarization effects.³⁴ It is by now well-recognized that standard Gaussian basis sets are in general not able to model properties which depend on the core electron density at sufficient accuracy. The problem is that such basis sets are optimized to describe the valence region and do not fulfill the correct nuclear cusp condition leading to slow convergence with increasing basis set size. The combination of slow convergence both with respect to the level of sophistication in theory and basis sets size leads to severe restrictions in the complexity of the studied systems. For systems with d-block metals, only small (often linear) systems have been treated comprehensively^{35,36} and with *ad hoc* devised basis sets. With the introduction of density functional theory, large parts of the dynamic electron correlation can be handled at much reduced cost, and calculation of hyperfine coupling tensors for d-block metal complexes is now tractable. Results from the p-block elements using DFT for hyperfine coupling constants have generally been promising,^{37–40} and several benchmark studies for d-block metal compounds have already emerged.^{36,41,42} Nevertheless, these benchmark studies focus mainly on the exchange-correlation functional, while the first-principle calculation of hyperfine coupling constants can be done at reduced cost and much more consistently if flexible, yet efficient basis sets were devised for the d-block metals. Several authors have shown that it is possible to modify Gaussian basis sets to describe core properties,^{23,34,43–49} but this has mostly focused on s- and p-block elements and few choices are available for the d-block.⁵⁰

In this paper, we extend the core-property basis aug-cc-pVTZ-J^{51–57} to the 3d metals Sc–Zn, such that accurate hyperfine couplings can be obtained for both metal and ligand spheres, using the same type of basis sets. This has not yet been possible within theoretical/computational d-block chemistry. It should be emphasized that the aug-cc-pVTZ-J basis sets for s- and p-block elements originally were optimized to describe nuclear magnetic

resonance spin–spin couplings. However, the d-block metals often contain unpaired electrons, usually situated in orbitals of high d character and have therefore traditionally been investigated by EPR. Seeing that the basis set requirements are similar in the description of NMR and EPR hyperfine couplings, we have here chosen an EPR framework, despite it meaning we have to deal with open-shell systems, which traditionally have been much more difficult to handle in terms of convergence and computer time.

METHODS

All calculations of hyperfine coupling constants were performed with the ORCA program⁵⁸ at the DFT level of theory. Throughout all calculations, the integration grid was always kept very large (IntAcc = 6 and AngularGrid = 7) to ensure that the core density was correctly described. Several studies on both main group^{59–61} and d-block metal complexes^{36,41,42} with a variety of different functionals have proven that A_{tot} can be quite sensitive to the exchange-correlation functional. Accordingly, we have used a small selection of functionals. For the calculation of hyperfine couplings, two GGA functionals, BP86^{62,63} and PBE,⁶⁴ a meta-GGA functional, TPSS,⁶⁵ and one hybrid-meta GGA functional, TPSSH,⁶⁵ were employed.

Basis Set Construction. The modification of the original aug-cc-pVTZ^{66–69} basis sets is performed for a set of molecules with atoms from the first row d-block metals (Figure 1, described in detail in the following subsection). This set is henceforth referred to as “training set 1”. Using this training set, the original basis sets are gradually decontracted. After decontraction, we add tight functions in an even-tempered manner as described in earlier work,^{51–57} starting with s-type functions. From the fully decontracted basis set, augmented with a sufficient number of tight s-type functions, we create two series where the first has tight p-type functions and the second, tight d-type functions. The tight p- and d-type functions which have a non-negligible effect on the hyperfine coupling tensor are included in the final basis sets. The uncontracted primitives and additional tight functions are recontracted using the molecular orbital coefficients of metal s-, p-, and d-character for the molecules in training set 1. To avoid bias in the final basis sets, the MO coefficients are compared to MO coefficients from a second set of molecules, henceforth denoted “training set 2” (Figure 2). Care has been taken to include metal complexes with very distinct ligand environments in the two training sets. To compare the MO coefficients from orbitals in the reference compounds (compounds in training set 1) and the corresponding metal complex in training set 2, the MOs must be renormalized. This is necessary as the orbitals with metal character also might contain some “contamination” of ligand orbitals, which will differ between training sets 1 and 2. Using the orbital of 1s character in TiF_3 as an

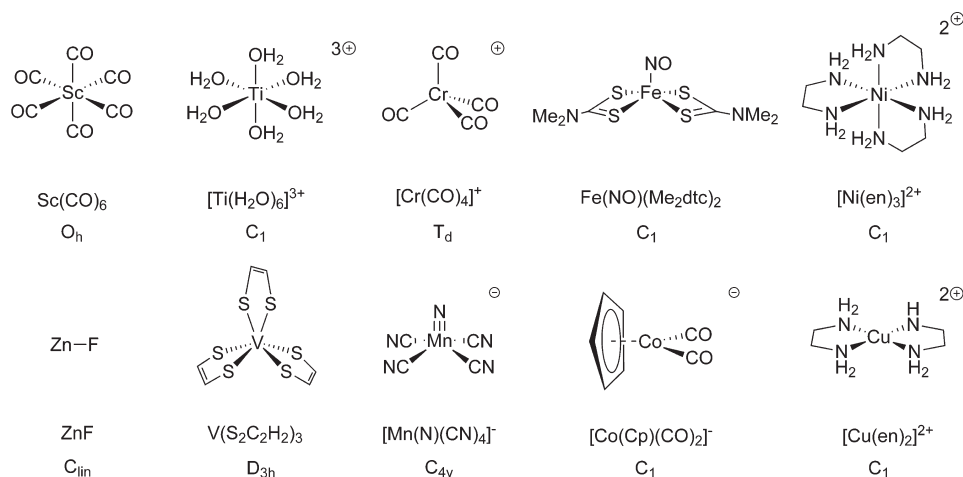


Figure 2. Training set 2 used under recontraction of the aug-cc-pVTZ-Juc to the aug-cc-pVTZ-J basis set.

example, the renormalization is done by using the ratio between the largest MO coefficient within the *s* primitive functions, comprising the 1s metal orbital and the corresponding metal *s* primitive MO coefficient in training set 2. All MO coefficients for *s*-type primitives within the orbital of metal 1s character in the reference compound are multiplied by this ratio. This method is repeated for orbitals of 2s and 3s, 2p and 3p, and 3d character. The renormalized coefficients are shown in Figures 9, 10, and 11. Despite this procedure being necessary to allow for a strict comparison, it has in fact little practical implication, as the original *s*-, *p*-, and *d*-type function coefficients and the renormalized coefficients are very close to each other.

For the 2p and 3p orbitals, one must choose between three orbitals, p_x , p_y , or p_z . A similar problem occurs in the contraction of the five d orbitals. Generally, the orbital with the largest p or d character is chosen, deemed from the MO coefficient sum of squares. For the p orbitals, this makes little difference, as the sum of squares are similar for all three orbitals and in all cases the p_z orbital is chosen. For the d orbitals, there is some difference in the sum of squares, and these have been investigated in detail.

Molecular Training Sets. The molecular geometries used are equilibrium geometries, optimized using the Gaussian program⁷⁰ with the BP86 functional and a TZVP basis set from the Ahlrichs group.^{71,72} For all molecules in the two training sets, very tight SCF convergence criteria and an ultrafine grid were used under optimization. Second-order derivatives were calculated analytically, and the absence of imaginary frequencies was used to confirm the minimum character of the optimized structures. Three requirements have been used in the selection of molecules for training set 1 and training set 2. First, the molecules should be experimentally well-established and preferably also have been investigated experimentally using EPR techniques. Second, the training set as a whole should be varied and represent a wide range of different ligands and coordination environments. Third, the molecules comprising training sets 1 and 2 should be computationally feasible. The last requirement naturally means that compromises between the two other requirements must be made. The scandium and zinc atoms imposed a particular challenge since they most often are found with empty or full d shells, respectively. Accordingly, they do not often show EPR activity, unless bounded to a radical ligand. Here, $\text{Sc}(\text{CN})_2$ and $\text{Zn}(\text{Me}_2\text{N}_2\text{C}_2\text{H}_2)\text{Cl}$ are used in training set 1. $\text{Sc}(\text{CN})_2$ is a rare d^1 EPR-active scandium system and has been identified using EPR in the gas phase. $\text{Zn}(\text{N}_2\text{C}_2\text{Me}_2)\text{Cl}$ is a model of the complex $\text{Zn}(\text{tBu}_2\text{N}_2\text{C}_2\text{H}_2)\text{Cl}$ which has been

shown to be EPR-active⁷³ due to a radical ligand. In training set 2, ZnF has been identified in the gas phase⁷⁴ and in a frozen neon matrix,⁷⁵ and the Zn hyperfine coupling tensor has been determined both experimentally⁷⁴ and theoretically.^{76,77} $\text{Sc}(\text{CO})_6$ has not been observed but is used here to represent low-valent organometallic compounds of the early d-block. TiF_3 has previously been investigated with DFT by several groups^{36,41,42,78,79} and experimentally by DeVore and Weltner⁸⁰ (EPR) and Hastie et al.⁸¹ (IR). The spin Hamilton parameters of the corresponding titanium complex in training set 2, $[\text{Ti}(\text{H}_2\text{O})_6]^{3+}$, have also been investigated experimentally,⁸² although it was not possible to resolve the hyperfine coupling to the titanium center. The structure of $[\text{Ti}(\text{H}_2\text{O})_6]^{3+}$ is known from X-ray crystallography on Ti^{3+} -doped host lattices.⁸³ The oxo complexes of vanadium and chromium in training set 1 were chosen, as they are biologically and industrially important while simultaneously being representative for the coordination chemistry of early d-block metals.⁸⁴ Structural^{85,86} (X-ray crystallography) and magnetic data,^{87,88} including EPR, exist for both compounds. Further, the EPR parameters of $\text{V}(\text{O})(\text{acac})_2$ have been investigated with DFT by Saladino and Larsen⁸⁹ and by Neese.²⁰ EPR and crystallographic data also exist for $\text{V}(\text{S}_2\text{C}_2\text{H}_2)_3$ ⁹⁰ in training set 2, whereas the chromium compound in this training set, $[\text{Cr}(\text{CO})_4]^+$, has only been identified by EPR in a solid krypton matrix.⁹¹ Likewise, MnO_3 has been matrix isolated and identified by EPR,⁹² but no other structural data exist. This molecule is known to be a difficult case and was also included in the training sets of Munzarová and Kaupp^{36,93} and by Kossmann et al.⁴¹ in their large studies of DFT functionals and their performance in prediction of hyperfine coupling parameters. It should be pointed out that we use C_{3v} symmetry for MnO_3 contrary to other studies,^{36,41,93} which use D_{3h} symmetry. With the BP86 functional, the D_{3h} geometry resulted in a structure with one imaginary frequency, corresponding to a bending motion toward C_{3v} , while the frequency analysis shows only real and positive frequencies for the C_{3v} equilibrium structure. For $[\text{Mn}(\text{N})(\text{CN})_4]^-$, which is used in test 2, experimental hyperfine couplings have also been determined.⁹⁴ The EPR parameters of $[\text{Mn}(\text{N})(\text{CN})_4]^-$ have further been investigated by DFT.^{36,42} With respect to the iron compounds in training sets 1 and 2, *trans*- $\text{Fe}(\text{en})_2\text{Cl}_2$ has not been isolated. It is used as a model for $\text{Fe}(\text{tmen})_2\text{Cl}_2$, which has been crystallographically characterized.⁹⁵ In the model, we use the same spin state ($S = 3/2$) as for the original

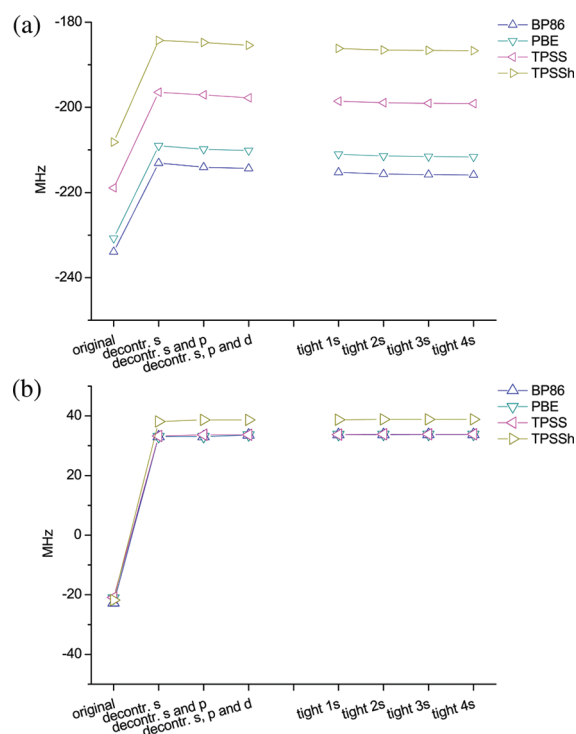


Figure 3. The effect on A_{iso} of stepwise decontraction of s-, p-, and d-type functions and further addition of four tight s-type functions for (a) TiF_3 and (b) $[Cr(O)Cl_4]^-$.

complex. The structure⁹⁶ and EPR parameters⁹⁷ of $Fe(NO)-(Me_2dtc)_2$ in training set 2 are known, and the latter have also been subjected to theoretical investigation.⁹⁸ The complexes $Co(acacen)$ and $[Co(Cp)(CO)_2]^-$ have both been investigated using EPR^{99–101} and for $Co(acacen)$ also using X-ray crystallography.¹⁰² Theoretical investigations of EPR parameters for $Co(acacen)$ and $[Co(Cp)(CO)_2]^-$ have also been conducted.^{103–105} Turning to the nickel compounds, $[Ni(mnt)_2]^-$ (training set 1) has undergone intense investigation, and both structural¹⁰⁶ and EPR data^{107–109} are known. Theoretical data on EPR parameters also exist from several sources.^{41,110,111} The corresponding complex in training set 2 is $[Ni(en)_3]^{2+}$. This complex has been structurally characterized,¹¹² but no experimental hyperfine coupling parameters exist. Finally, the two copper complexes $Cu(CO)_3$ and $[Cu(en)_2]^{2+}$ both exist. EPR data, including Cu hyperfine coupling tensors, have been measured.^{113–115} $[Cu(en)_2]^{2+}$ has also been studied by X-ray crystallography.¹¹⁶ The two compounds have also been used for DFT validation studies of EPR hyperfine coupling tensors.^{36,41,42,117}

Uncontraction and Extension of the Basis Sets. For the molecules in training set 1, the metal hyperfine coupling constants were calculated with the original and uncontracted aug-cc-pVTZ basis sets. For TiF_3 , $[Cr(O)Cl_4]^-$, $trans-Fe(en)_2Cl_2$, and $[Ni(mnt)_2]^-$ molecules, the decontraction was performed in three steps, such that s-type functions were decontracted first, followed by p-type functions, and d-type functions were decontracted last. The results are shown in Figures 3 and 4. For the first coordination sphere (the atoms directly coordinated to the metal center), the same type of basis set (including the level of decontraction) as for the metal was used. Atoms not directly attached to the metal were always described with the aug-cc-pVDZ basis without any modifications. Under the addition of tight s-type functions, tight functions were added on

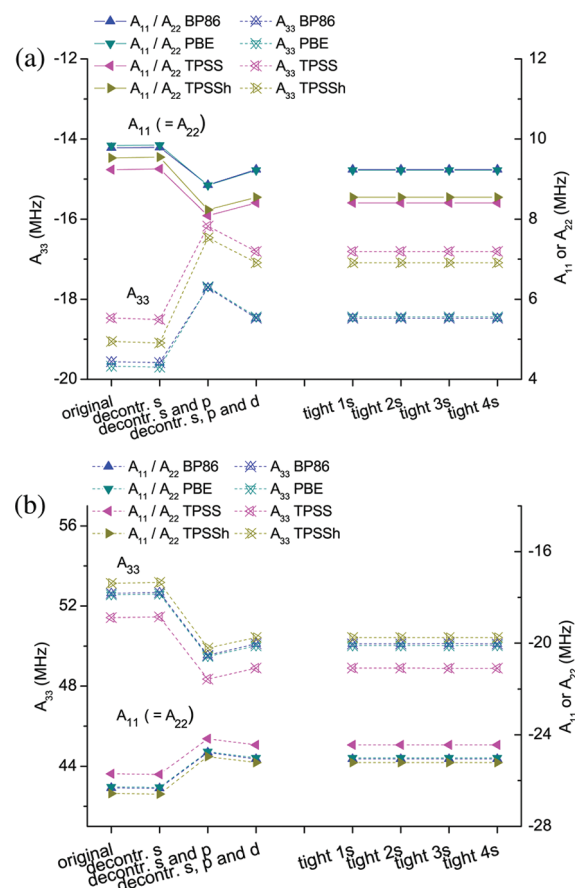


Figure 4. The effect on A_{SD} of stepwise decontraction of s-, p-, and d-type functions and further addition of four tight s-type functions for (a) TiF_3 and (b) $[Cr(O)Cl_4]^-$.

both the metal and all atoms within the first coordination sphere, such that the number of added s-type functions always is the same for metal and the first coordination sphere. This was done in order to keep the basis set balanced around the metal center but allows also to investigate whether the aug-cc-pVTZ-J basis sets are adequate to describe superhyperfine couplings, i.e., couplings to ligand atoms. The result from these calculations will appear in a subsequent publication.

Basis Set Recontraction. The calculations used for the basis set reconstructions were done with the same specifications as under the decontraction. The only exception was the use of the original aug-cc-pVTZ basis for the first coordination sphere instead of a fully decontracted basis with added tight functions. We have in parallel calculations shown that modifications of the basis sets in the ligand sphere appear to have only a minor influence on the metal coupling constant.

RESULTS AND DISCUSSION

Basis Set Convergence. For the compounds TiF_3 , $[Cr(O)Cl_4]^-$, $trans-Fe(en)_2Cl_2$, and $[Ni(mnt)_2]^-$ the effect on A_{iso} and A_{SD} due to stepwise decontraction of s-, p-, and d-type functions was investigated in detail. The three components in the 3×3 spin-dipolar hyperfine coupling tensor (in the diagonal representation) are here denoted A_{11} , A_{22} , and A_{33} . Due to symmetry, two of these components are equivalent in TiF_3 and $[Cr(O)Cl_4]^-$ (axial symmetry), and generally

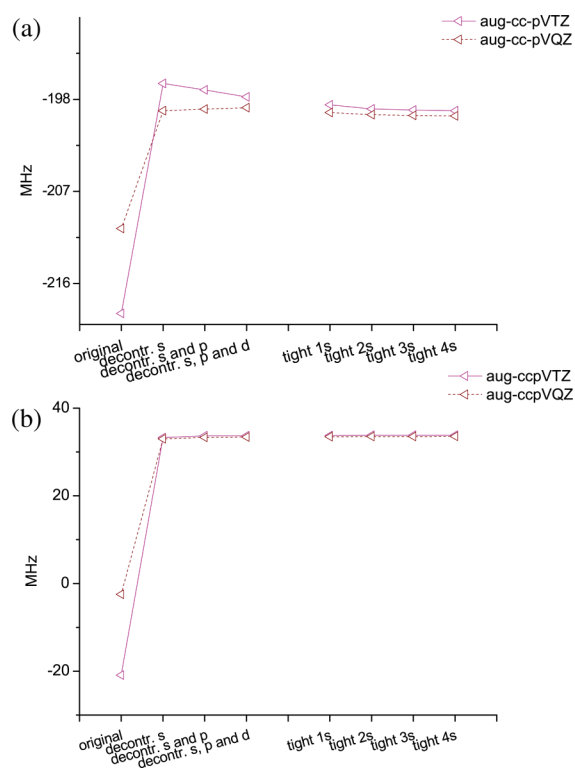


Figure 5. The effect on A_{iso} of stepwise decontraction of s-, p-, and d-type functions and further addition of four tight s-type functions using aug-cc-pVTZ and aug-cc-pVQZ basis sets for (a) TiF_3 and (b) $[\text{Cr}(\text{O})\text{Cl}_4]^-$ (TPSS results).

the figures will be restricted to contain only these two compounds, as they are the most simple. The conclusions can be generalized to include *trans*- $\text{Fe}(\text{en})_2\text{Cl}_2$ and $[\text{Ni}(\text{mnt})_2]^-$. The convergence behavior for the isotropic and spin-dipolar coupling tensors are shown in Figures 3 and 4, respectively. For comparison, the addition of tight s-type functions is also included in these two figures. Note also that in the figure with spin dipolar tensors (Figure 4) two axes have been used, as the individual tensor components can obtain quite different values. However, the same scale is used on both axes.

The decontraction of s-type functions constitutes by far the largest contribution to A_{iso} . Decontraction of p- and d-type functions has sometimes a small effect (usually below 1 MHz). Since only s-type functions contribute directly to the spin density on the nucleus, this can be ascribed to a slightly different screening effect on the (uncontracted) s-type functions from the basis set left with contracted p- or d-type functions as compared to the basis set where these functions are uncontracted. As expected from the detailed work by Kaupp and Munzarová on spin polarization mechanisms,⁹³ decontraction of p- and d-type functions affects the dipolar term (Figure 4) and is thus generally not negligible for precise calculations. The effect on A_{iso} from decontraction of the individual s-type functions is presumably similar for the compounds $\text{Sc}(\text{CN})_2$, $\text{V}(\text{O})(\text{acac})_2$, MnO_3 , $\text{Co}(\text{acac})$, $\text{Cu}(\text{CO})_3$, and $\text{Zn}(\text{Me}_2\text{N}_2\text{C}_2\text{H}_2)\text{Cl}$, for which the decontraction was performed in one step.

The addition of tight s-type functions is less important than the decontraction of s-type functions, but in order to saturate the basis set fully with respect to A_{iso} , the addition of tight functions is necessary. Although the basis sets for some of the metals in training set **1** are saturated with a lower number, four tight s-type

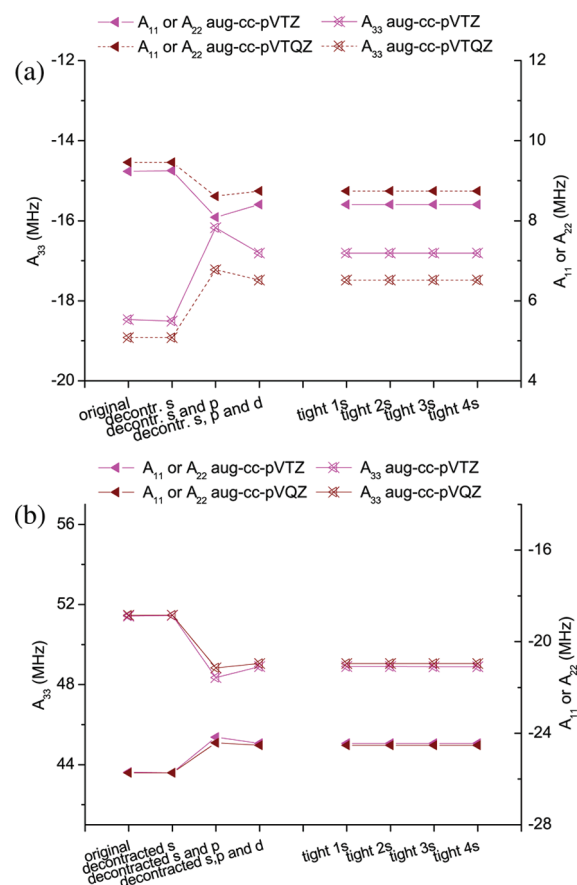


Figure 6. The effect on A_{SD} of stepwise decontraction of s-, p-, and d-type functions and further addition of four tight s-type functions using aug-cc-pVTZ and aug-cc-pVQZ basis sets for (a) TiF_3 and (b) $[\text{Cr}(\text{O})\text{Cl}_4]^-$ (TPSS results).

functions have been added over the whole series for consistency. This is also the number of tight s-type functions used in the construction of aug-cc-pVTZ-J basis sets for p-block elements.^{51–57}

The addition of four tight p-type functions and four tight d-type functions was not necessary, but inclusion of the first tight p-type (or d-type) function sometimes shows a small effect. We have chosen to include therefore one tight p- and one tight d-type function in the final, uncontracted basis sets (see tables in the Supporting Information). These are here denoted aug-cc-pVTZ-Juc in accordance with earlier work.^{51–57}

We have also considered the quadruple- ζ basis set aug-cc-pVQZ as the underlying basis set. The result of stepwise decontraction of this basis set is shown in Figure 5 for the isotropic part and in Figure 6 for the spin-dipolar part of the hyperfine coupling tensor. Although decontraction of the s-type functions has a smaller effect on A_{iso} than with aug-cc-pVTZ basis sets, it is still by far the most important step in the decontraction. p- and d-type functions are again primarily important for the spin-dipolar part. Regarding the addition of s-type functions, three tight s-type functions are included, which again leads to basis sets similar to the main group atoms.^{51–57} The addition of tight p- and d-type functions has little or no effect on A_{iso} and A_{SD} . As was the case in the aug-cc-pVTZ-Juc basis sets, one tight p- and one d-type function has been included in the final aug-cc-pVQZ-Juc basis sets. The aug-cc-pVQZ-Juc basis sets will not be considered

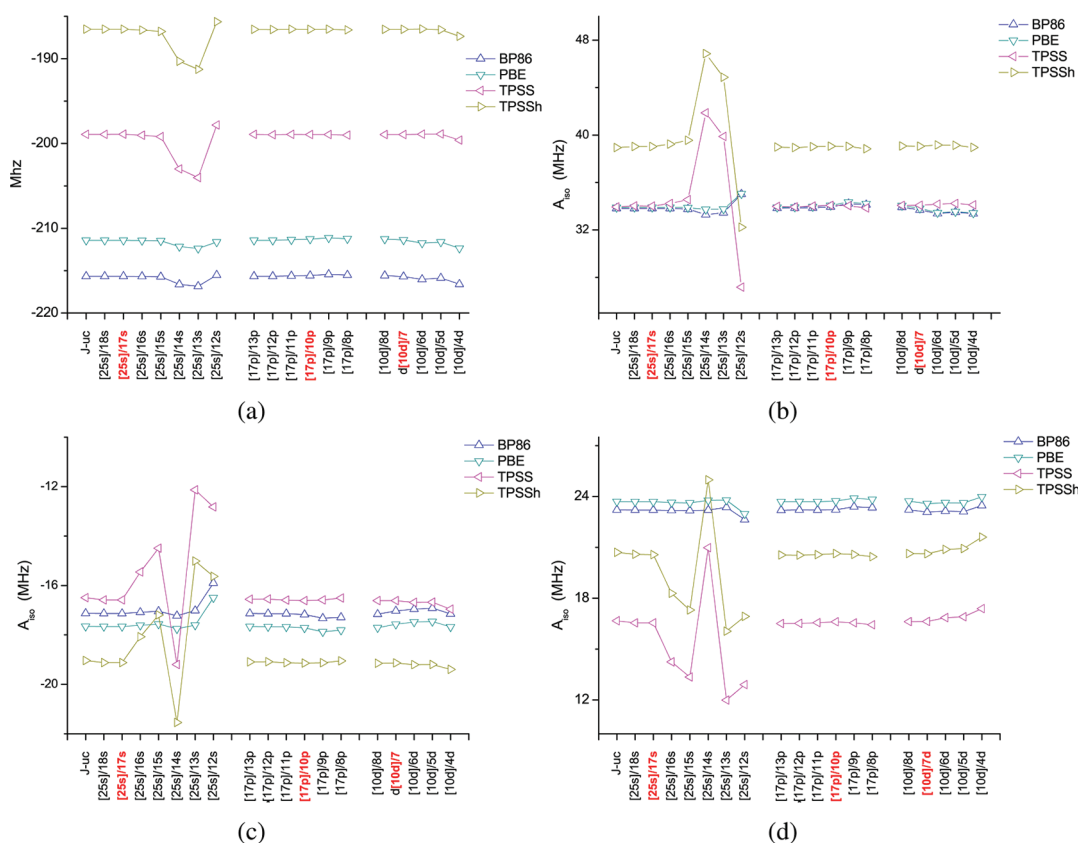


Figure 7. The effect on A_{iso} upon recontraction for (a) TiF_3 , (b) $[\text{Cr}(\text{O})\text{Cl}_4]^-$, (c) $\text{trans-Fe}(\text{en})_2\text{Cl}_2$, and (d) $[\text{Ni}(\text{mnt})_2]^-$. The red labels indicate the chosen recontraction scheme for the final basis sets.

further here, as they become too large for most purposes and the gained accuracy compared to the aug-cc-pVTZ-Juc basis sets is quite small, c.f., Figures 5 and 6.

Basis Set Recontraction. In order to reduce the size of the aug-cc-pVTZ-Juc basis set and to use it consistently with the previously developed aug-cc-pVTZ-J basis sets for lighter atoms, the uncontracted basis sets must be recontracted. As described previously,^{51–57} we use MO coefficients from a training set of molecules as contraction coefficients. In this work, the molecules in training set 1 are used. The recontraction has been done for the s-, p-, and d-type functions, separately with careful consideration of the effect on the coupling constant. This is shown for A_{iso} in Figure 7. Contrary to the decontraction study in the previous section, the investigation of how strongly the basis sets could be recontracted was performed for all molecules in training sets 1 and 2. This more extensive investigation was carried out since the effect of recontraction on A_{iso} was found to be very sensitive to both the used functional and the metal in question. Figure 7 shows a representative set of scenarios for four different molecules containing the metals Ti, Cr, Fe, and Ni. It can be seen that by recontracting the s-type functions from (25s) to [17s], using a general contraction scheme where the first 11 primitive s functions (with largest exponents) are contracted to three functions, all metals keep an accurate s-type function basis set. This contracted basis set is still converged with respect to A_{iso} , also for the metals not shown here (see the Supporting Information). More primitive s-type functions could have been included in the contracted functions for Ti and Cr (and some of the other metals,

as shown in the Supporting Information), but the chosen contraction scheme is preferred since it is consistent over the d-block.

The effect on A_{SD} is much more uniform and not nearly as dependent on functional and metal as A_{iso} (Figure 8). The focus is therefore in the following only on TiF_3 and $[\text{Cr}(\text{O})\text{Cl}_4]^-$, as these compounds have again the simplest spin dipolar tensors, while still being representative for all of the metals. A_{SD} can be described with good accuracy with a (17p) to [10p] general contraction, where the 10 primitives with highest exponents have been contracted to two functions. Finally, the d-type functions are described by a (10d) to [7d] general contraction, with the four steepest primitive d functions contracted to one function.

The method of using MO coefficients from a training set as basis set contraction coefficients has been criticized for leading to final basis sets with a bias toward the molecules within the training set.⁴⁷ Despite this concern being reasonable, it will be shown here by comparing coefficients obtained with training sets 1 and 2 that such bias is of little or no importance for the final basis sets. As an illustration, the molecular orbital coefficient of the orbitals with 1s, 2s, and 3s character are shown for TiF_3 in Figure 9.

Included in this figure are also the orbitals of 1s, 2s, and 3s character for $[\text{Ti}(\text{H}_2\text{O})_6]^{3+}$. Obviously, the compound $[\text{Ti}(\text{H}_2\text{O})_6]^{3+}$ has both different coordination geometry and also different ligands, compared to the reference compound (TiF_3). However, the titanium s-type function coefficients in TiF_3 and $[\text{Ti}(\text{H}_2\text{O})_6]^{3+}$ are very similar. This is not very surprising for the inner orbitals 1s and 2s, but even for the 3s orbital, which lies just below the valence orbitals, the coefficients are still almost identical. Figure 10 compares the TiF_3 and $[\text{Ti}(\text{H}_2\text{O})_6]^{3+}$ MO coefficients for the

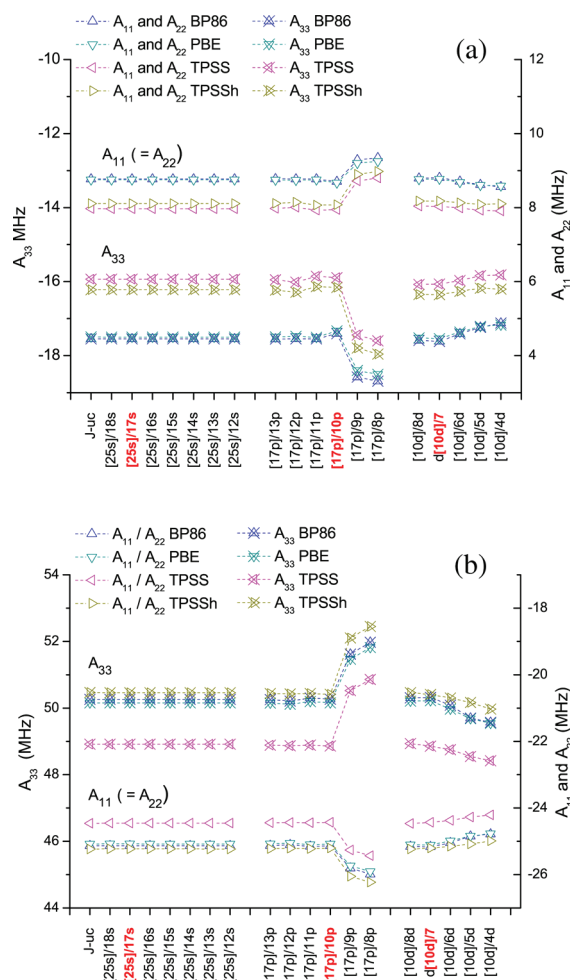


Figure 8. The effect on A_{SD} upon reconstruction for (a) TiF_3 and (b) $[\text{Cr}(\text{O})\text{Cl}_4]^-$. The red labels indicate the chosen reconstruction scheme for the final basis sets.

primitive Ti p_z orbitals in the MOs of 2p and 3p character on titanium. The same conclusion as for the s-type functions is reached: For the 2p and even for the 3p orbital, the coefficients are almost identical and thus apparently rather independent of the compound. Closer analysis shows, however, that differences in MO coefficients are larger for the more diffuse orbitals, while primitives with large exponents always have almost identical coefficients. Only the d orbitals show slightly different MO coefficients for the primitives (see Figure 11). But the d orbitals are in general more difficult, as they comprise the valence electrons. Accordingly, the d shell is usually partly filled, and it is not always possible to use the same d orbital in the reference compound (from training set 1) as in the corresponding compound in training set 2. Therefore, only the first four primitives are included in the contraction, as the MO coefficients for these are almost identical.

We have carefully tested all other molecules in training sets 1 and 2 in this manner and reach the same conclusions as for TiF_3 and $[\text{Ti}(\text{H}_2\text{O})]^{3+}$. With the great span of metals and different ligand spheres in these two training sets, we thus conclude that the method of using MO coefficients as reconstruction coefficients leads to basis sets with little or no bias. As a final comment on this matter, we did also perform a calculation with the Mn^{2+} ion. This calculations showed the coefficients to vary little between mole-

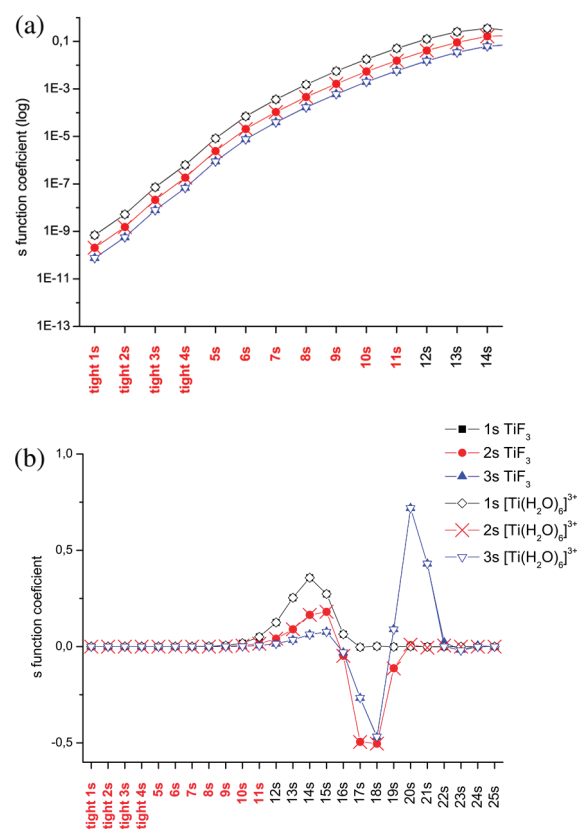


Figure 9. (a) The first 14 MO coefficients for Ti s-type primitives in the orbitals with 1s, 2s, and 3s character for Ti (logarithmic scale). (b) The full range of all 25 s-type primitives where the first four are the added tight functions and the remaining 21 from the original aug-cc-pVTZ basis set (TPSS results). The primitive functions with red labels are the primitives used in the reconstructions.

cules and atoms, but to establish this relationship more firmly, more extensive calculations using several ions or neutral atoms would be needed.

Dependence on Exchange-Correlation Functional. *Contraction Coefficients.* For the compounds TiF_3 , $[\text{Cr}(\text{O})\text{Cl}_4]^-$, $\text{trans-Fe}(\text{en})_2\text{Cl}_2$, and $[\text{Ni}(\text{mnt})_2]^-$, the effect of using different functionals on the values of the contraction coefficients was investigated in some detail. In line with other studies on main group atoms,^{47,57} this turned out to be of little importance. As an example, the MO coefficients for the titanium p- and d-type primitives in orbitals of 3p and 3d character are shown in Figure 12. The same pattern as in Figure 12 is seen for the other three compounds.

Hyperfine Coupling Constants. It is noted by comparing Figure 3a and b that the metal isotropic hyperfine coupling tensor is very sensitive to the used functional, although this dependence again depends on the metal in question. Thus, a much wider span in the final hyperfine coupling constants arises from the four applied functionals for TiF_3 , compared to $[\text{Cr}(\text{O})\text{Cl}_4]^-$. As can be seen from the tables collected in the Supporting Information, the span in the final A_{iso} values for the four different functionals is very different across the series of 3d metals. However, the convergence behavior in the basis sets is always completely similar for all of the functionals used. This is also the case for the dipolar couplings, although the final values in A_{SD} depend much less critically on the applied functional (here, the spans are within 2 MHz).

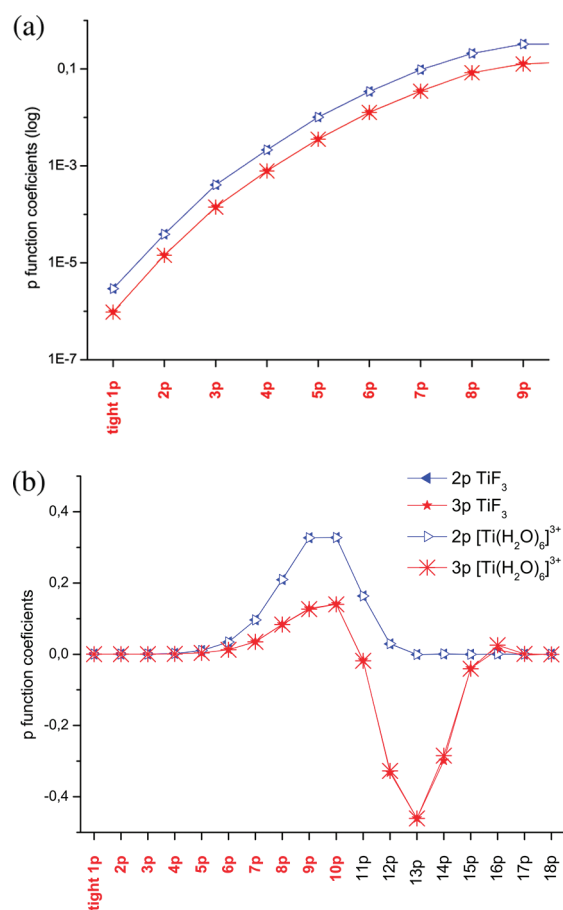


Figure 10. (a) The first nine MO coefficients for titanium p-type primitives in the orbitals of titanium $2p_z$ and $3p_z$ character (logarithmic scale). (b) The full range of all 17 p-type primitives where the first is the added tight function and the remaining 16 are from the original aug-cc-pVTZ basis set (TPSS results). The primitive functions with red labels are the primitives used in the reconstructions.

In the case of MnO_3 , inclusion of exact exchange occasionally leads to complete deterioration of the results, and for this special case, hybrid functionals are not recommended. The same conclusion was reached by Kaupp and Munzarova⁹³ even though they used a different equilibrium geometry (D_{3h}). The equilibrium geometry itself of this molecule is also very dependent on the used functional; where the hybrid functional B3LYP leads to an equilibrium structure of D_{3h} symmetry, the BP86 functional used here leads instead to a C_{3v} equilibrium geometry.

The difference in the effect of recontraction upon both A_{iso} and A_{SD} is also quite pronounced. Here, *meta*-GGA functionals are found to be most sensitive to the level of recontraction, and a much more conservative choice of recontraction scheme must be chosen when such functionals are used. Detailed investigation of the underlying mechanisms are currently being undertaken.

COMPARISON WITH EXPERIMENT

A comparison of measured hyperfine coupling constants with results of calculations at equilibrium geometries is not trivial, as both environmental effects from solvation and ro-vibrational contributions can have substantial effects on the calculated values. Furthermore, for transition metals also relativistic effects can be

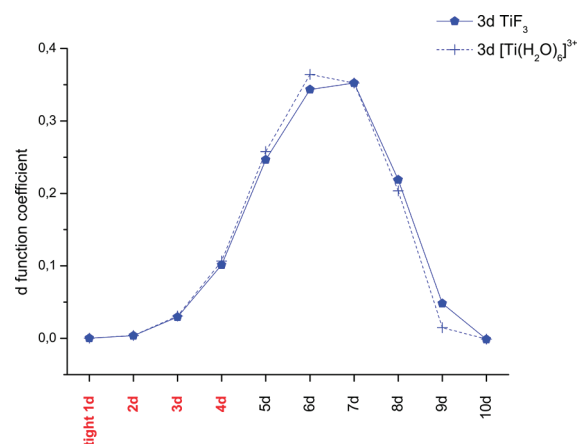


Figure 11. The 10 MO coefficients for titanium d-type primitives in the orbitals of titanium $3d_{z^2}$ character. The first function is the added tight function, and the remaining nine are from the original aug-cc-pVTZ basis set (TPSS results). The primitive functions with red labels are the primitives used in the reconstructions.

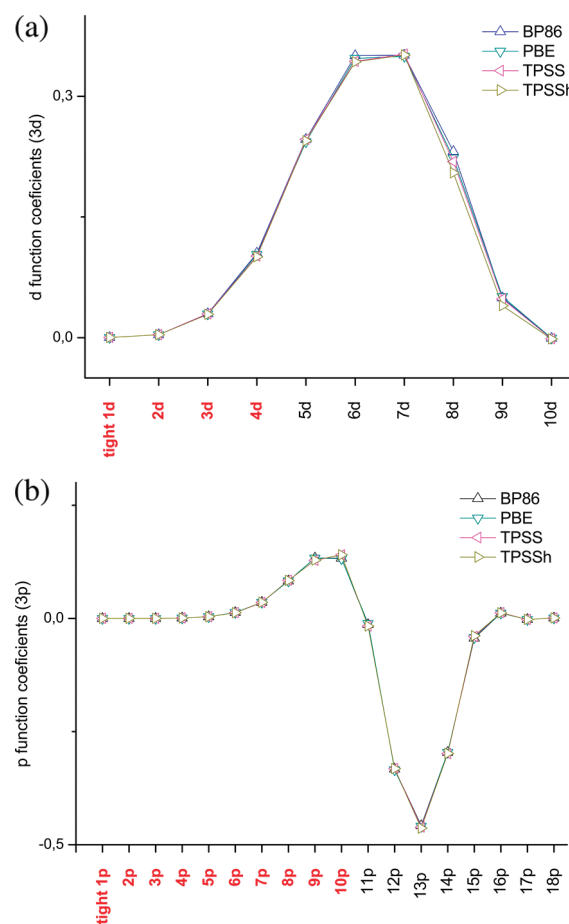


Figure 12. MO coefficients for Ti p- and d-type primitives in the orbitals of $3p$ (a) and $3d$ (b) character for TiF_3 . The primitive functions with red labels are the primitives used in the reconstructions.

important. Nevertheless, some of the compounds from training set I, where the hyperfine coupling has been resolved from the experimental EPR spectrum, are here compared to the calculated

Table 1. Calculated versus Experimental Values for Hyperfine Coupling Tensors (TPSS Results)

molecule	basis	A_{11}^{SD}	A_{22}^{SD}	A_{33}^{SD}	A_{iso}
TiF ₃	aug-cc-pVTZ	9.23	9.23	-18.47	-218.93
	aug-cc-pCVTZ	8.61	8.61	-17.21	-199.13
	aug-cc-pVTZ-J	8.03	8.03	-16.07	-198.95
exptl. ⁸⁰		6.60	6.60	-12.40	-184.8
V(O)(acac) ₂	aug-cc-pVTZ	94.49	108.24	-202.73	26.82
	aug-cc-pCVTZ	104.13	91.13	-195.3	-144.87
	aug-cc-pVTZ-J	104.21	92.03	-196.24	-219.26
exptl. ⁸⁷		116	112	-228	-307
MnO ₃	aug-cc-pVTZ	-18.87	-18.87	37.73	-124.16
	aug-cc-pCVTZ	-90.59	-90.58	181.17	1818.62
	aug-cc-pVTZ-J	-92.83	-92.83	185.66	1754.11
exptl. ⁹²		-80.00	-80.00	160.00	1612.00
Cu(CO) ₃	aug-cc-pVTZ	-57.21	-57.21	114.43	-18.62
	aug-cc-pCVTZ	-47.91	-47.91	117.50	-9.30
	aug-cc-pVTZ-J	-59.25	-59.25	118.50	2.67
exptl. ¹¹³		-81.17	-81.17	162.33	71.17

hyperfine couplings in Table 1 (see also Supporting Information). The original aug-cc-pVTZ and the core-correlation basis set aug-cc-pCVTZ are also included for comparison. Although the aug-cc-pCVTZ basis sets lead to some improvement over the original aug-cc-pVTZ basis sets, the results are still quite far from the converged aug-cc-pVTZ-J basis set results (all results in Table 1 are obtained with the TPSS functional). For most of the shown examples, reasonable qualitative agreement is obtained between experimental hyperfine coupling tensors and the aug-cc-pVTZ-J basis set results. Nevertheless, it is clear that more work is needed to obtain quantitative agreement between theory and experiment.

The compound MnO₃ is of some interest, since the functional used for geometry optimization (BP86) yielded a C_{3v} structure, rather than the experimentally observed D_{3h} structure, while the B3LYP functional leads to the D_{3h} structure. However, the experimental structure is deduced from the size of the ⁵⁵Mn hyperfine coupling alone, and the studies presented here show that couplings of similar magnitude are calculated from a C_{3v} equilibrium structure. A detailed study of the MnO₃ equilibrium structure is beyond the scope of this work, but these observations definitely call for further investigations.

CONCLUSIONS

The aug-cc-pVTZ-J series of basis sets has been extended to the 3d metals Sc–Zn. First, the convergence of the Fermi-contact and spin-dipolar terms (A_{iso} and A_{SD}) was investigated with respect to decontracting the s-, p-, and d-type functions in the original aug-cc-pVTZ basis sets and then further with respect to the addition of tight s-, p-, and d-type functions. As discovered for the main group elements, the original aug-cc-pVTZ basis sets are too contracted in the s shells to describe core properties.¹¹⁸ The training set used in the generation of the basis sets consisted mainly of experimentally well-known compounds, which often also have been studied using EPR techniques.

In order to prevent bias toward a particular exchange-correlation functional, we have employed four different functionals. Although there were found to be quite substantial differences in the final predicted values of especially A_{iso} , the convergence of the results with respect to decontraction of the basis sets and the

addition of tight functions was generally not dependent on the employed functional. A modification of aug-cc-pVTZ (uncontracted) with four added tight s-type functions, one tight p-, and one tight d-type function was found to give a hyperfine coupling tensor which was saturated with respect to the basis set. This basis set is denoted aug-cc-pVTZ-Juc, which comprises a 25s17p10d3f2g set of functions. These uncontracted basis sets have been recontracted to the final aug-cc-pVTZ-J basis sets according to (25s17p10d3f2g)/[17s10p7d3f2g] and using a general contraction scheme. MO coefficients from the molecules in a training set (training set 1) were used as contraction coefficients. By careful comparison with a second training set (training set 2), it could be shown that a possible bias toward the molecules in training set 1 largely was avoided. Furthermore, the obtained coefficients have been shown to be insensitive to the functional used. Future work will be directed to benchmark the final aug-cc-pVTZ-J basis sets against a larger compilation of systems with experimentally known hyperfine coupling constants employing a larger set of functionals and also high-level *ab initio* methods.

ASSOCIATED CONTENT

S Supporting Information. Detailed tables of the calculated hyperfine coupling constants (experimental data is included when available) at different contraction levels. With the exception of Figures 5 and 6, all figures have been constructed from these tables. The data from Figures 5 and 6 can be obtained from E.D.H. (edh@ifk.sdu.dk). This material is available free of charge via the Internet at <http://pubs.acs.org>.

AUTHOR INFORMATION

Corresponding Author

*E-mail: sauer@kiku.dk.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENT

This work has been supported by the Danish Center for Scientific Computing and the Danish Natural Science Research Council/Danish Councils for Independent Research. J.K. thanks the Lundbeck Foundation and S.P.A.S thanks the Carlsberg foundation for financial support.

REFERENCES

- (1) Kaupp, M.; Bühl, M.; Malkin, V. G. *Calculation of NMR and EPR Parameters. Theory and Applications*; Wiley: New York, 2004.
- (2) Bleaney, B.; Abragam, A. *Electron Paramagnetic Resonance of Transition Metal Ions*; Oxford University Press: Oxford, U. K., 1970.
- (3) Tuttle, T. R., Jr.; Ward, R. L.; Weissman, S. I. *J. Chem. Phys.* **1956**, *25*, 189–189.
- (4) Katz, T. J.; Strauss, H. L. *J. Chem. Phys.* **1960**, *32*, 1873–1875.
- (5) Fessenden, R. W.; Schuler, R. H. *J. Chem. Phys.* **1963**, *39*, 2147–2195.
- (6) Schäfer, K. O.; Bittl, R.; Lenzian, F.; Barynin, V.; Weyhermüller, K.; Wieghardt, T.; Lubitz, W. *J. Phys. Chem. B* **2003**, *107*, 1242–1250.
- (7) Veselov, A.; Sun, H.; Sienkiewicz, A.; Taylor, H.; Burger, R. M.; Scholes, C. P. *J. Am. Chem. Soc.* **1995**, *117*, 7508–7512.
- (8) Veselov, A.; Burger, R. M.; Scholes, C. P. *J. Am. Chem. Soc.* **1998**, *120*, 1030–1033.
- (9) Pryce, M. L. H. *Phys. Rev.* **1950**, *80*, 1107–1108.

- (10) Bleaney, B.; Ingram, D. J. E. *Proc. Phys. Soc. A* **1950**, *63*, 409–411.
- (11) Abragam, A.; Pryce, M. L. H. *Proc. R. Soc. London, Ser. A* **1951**, *205*, 135–153.
- (12) Abragam, A.; Pryce, M. L. H. *Proc. R. Soc. London, Ser. A* **1951**, *206*, 173–191.
- (13) Abragam, A.; Horowitz, J.; Pryce, M. L. H.; Morton, K. W. *Proc. R. Soc. London, Ser. A* **1955**, *230*, 169–187.
- (14) Davydov, R.; Kappl, R.; Hüttermann, J.; Peterson, J. A. *FEBS Lett.* **1991**, *295*, 113–115.
- (15) Davydov, R.; Makris, T. M.; Kofman, V.; Werst, D. E.; Sligar, S. G.; Hoffman, B. M. *J. Am. Chem. Soc.* **2001**, *123*, 1403–1415.
- (16) Davydov, R.; Satterlee, J. D.; Fujii, H.; Sauer-Masarwa, A.; Busch, D. H.; Hoffman, B. M. *J. Am. Chem. Soc.* **2003**, *125*, 16340–16346.
- (17) Davydov, R.; Perera, R.; Jin, S.; Yang, T.; Bryson, T. A.; Sono, M.; Dawson, J. D.; Hoffman, B. M. *J. Am. Chem. Soc.* **2005**, *127*, 1403–1413.
- (18) Mukhopadhyay, S.; Mandal, S. K.; Bhaduri, S. S.; Armstrong, W. H. *Chem. Rev.* **2004**, *104*, 3981–4026.
- (19) Sauer, S. P. A. *Molecular Electromagnetism: A Computational Chemistry Approach*; Oxford University Press: Oxford, U. K., 2011.
- (20) Neese, F. *J. Chem. Phys.* **2003**, *118*, 3939–2948.
- (21) Watson, R. E.; Freeman, A. J. *Phys. Rev.* **1961**, *123*, 2027–2047.
- (22) Engels, B.; Peyerimhoff, S. D.; Karna, S. P.; Grein, F. *Chem. Phys. Lett.* **1988**, *152*, 397–401.
- (23) Engels, B.; Peyerimhoff, S. D. *J. Phys. B: At. Mol. Opt. Phys.* **1988**, *88*, 3459–3471.
- (24) Feller, D.; Davidson, E. R. *J. Chem. Phys.* **1988**, *88*, 7580–7587.
- (25) Bauschlicher, C. W., Jr.; Langhoff, S. R.; Partridge, H.; Chong, D. P. *J. Chem. Phys.* **1988**, *89*, 2985–2992.
- (26) Funken, K.; Engels, B.; Peyerimhoff, S. D.; F. Grein, F. *Chem. Phys. Lett.* **1990**, *172*, 180–186.
- (27) Chipman, D. M.; Carmichael, I.; Feller, D. *J. Phys. Chem.* **1991**, *95*, 4702–4708.
- (28) Engels, B. *Chem. Phys. Lett.* **1991**, *179*, 398–404.
- (29) Engels, B. *Theor. Chim. Acta* **1993**, *86*, 429–437.
- (30) Engels, B. *J. Chem. Phys.* **1994**, *100*, 1380–1386.
- (31) Perera, S. A.; Watt, J. D.; Bartlett, R. J. *J. Chem. Phys.* **1994**, *100*, 1425–2434.
- (32) Kong, J.; Boyd, R. J.; Erikson, L. A. *J. Chem. Phys.* **1995**, *102*, 3674–3678.
- (33) Engels, B. *Acta Chem. Scand.* **1997**, *51*, 199–210.
- (34) Chipman, D. M. *Phys. Rev.* **1989**, *39*, 475.
- (35) Knight, L. B., Jr.; Babb, R.; Ray, M.; Banisaukas, T. J., III; Russon, L.; Dailey, S.; Davidson, E. R. *J. Chem. Phys.* **1996**, *105*, 10237–10250.
- (36) Munzarová, M.; Kaupp, M. *J. Phys. Chem. A* **1999**, *103*, 9966–9983.
- (37) Ishii, N.; Shimizu, T. *Chem. Phys. Lett.* **1994**, *225*, 462–466.
- (38) Ishii, N.; Shimizu, T. *Chem. Phys. Lett.* **1995**, *235*, 614–616.
- (39) Barone, V. *Theor. Chem. Acc.* **1995**, *91*, 113–128.
- (40) Batra, R.; Giese, B.; Spichty, M.; Gescheidt, G.; Houk, K. N. *J. Phys. Chem.* **1996**, *100*, 18371–18379.
- (41) Kossmann, S.; Kirchner, B.; Neese, F. *Mol. Phys.* **2007**, *105*, 2049.
- (42) Abuznikov, A. V.; Kaupp, M.; Malkin, V. G.; Reviakine, R.; Malkina, O. L. *Phys. Chem. Chem. Phys.* **2002**, *4*, 5467–5474.
- (43) Chipman, D. M. *Theor. Chim. Acta* **1989**, *76*, 73–84.
- (44) Chipman, D. M. *J. Chem. Phys.* **1989**, *91*, 5455–5465.
- (45) Chipman, D. M. *Theor. Chim. Acta* **1992**, *82*, 93–115.
- (46) Barone, V. In *Recent Advances in Density Functional Methods Part I*; Chong, D. P., Ed.; World Scientific: Singapore, 1996.
- (47) Jensen, F. *J. Chem. Theory Comput.* **2006**, *2*, 1360–1369.
- (48) Barone, V.; Cimino, P.; Stendardo, E. *J. Chem. Theory Comput.* **2008**, *4*, 751–764.
- (49) Kjær, H.; Sauer, S. P. A. *J. Chem. Theory Comput.* **2011**, Online: <http://dx.doi.org/10.1021/ct200546q>.
- (50) Neese, F. *Inorg. Chim. Acta* **2002**, *337*, 181–192.
- (51) Enevoldsen, T.; Oddershede, J.; Sauer, S. P. A. *Theor. Chem. Acc.* **1998**, *100*, 275–284.
- (52) Sauer, S. P. A.; Raynes, W. T. *J. Chem. Phys.* **2000**, *113*, 3121–3129.
- (53) Sauer, S. P. A.; Raynes, W. T.; Nicholls, R. A. *J. Chem. Phys.* **2001**, *115*, 5994–6006.
- (54) Provasi, P. F.; Aucar, G. A.; Sauer, S. P. A. *J. Chem. Phys.* **2001**, *115*, 1324–1334.
- (55) Barone, V.; Provasi, P. F.; Peralta, J. E.; Snyder, J. P.; Sauer, S. P. A.; Contreras, R. H. *J. Phys. Chem. A* **2003**, *107*, 4748–4754.
- (56) Rusakov, Y. Y.; Krivdin, L. B.; Sauer, S. P. A.; Levanova, E. P.; Levkovskaya, G. G. *Magn. Reson. Chem.* **2010**, *48*, 633–637.
- (57) Provasi, P. F.; Sauer, S. P. A. *J. Chem. Phys.* **2010**, *133*, 054308.
- (58) Neese, F. ORCA - An ab initio, DFT and semiempirical Program package, Version 2.7. University of Bonn, 2009.
- (59) Erikson, L. A.; Wang, J.; Boyd, R. J. *Chem. Phys. Lett.* **1993**, *211*, 88–93.
- (60) Erikson, L. A.; Wang, J.; Boyd, R. J.; Lunell, S. J. *Phys. Chem.* **1994**, *98*, 792–799.
- (61) Erikson, L. A.; Malkina, O. L.; Malkin, V. G.; Salahub, D. R. *J. Chem. Phys.* **1994**, *100*, 5066–5075.
- (62) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648–5652.
- (63) Perdew, J. P. *Phys. Rev. B* **1986**, *33*, 8822–8824.
- (64) Perdew, J. P.; Burke, K.; Wang, Y. *Phys. Rev. Lett.* **1996**, *77*, 3865–3868. Erratum: **1997**, *78*, 1396.
- (65) Tao, J.; Perdew, J. P.; Staroverov, S. G. *Phys. Rev. Lett.* **2003**, *91*, 146401.
- (66) Dunning, T. H., Jr. *J. Chem. Phys.* **1989**, *90*, 1007–1023.
- (67) Kendall, R. A.; Dunning, T. H.; Harrison, R. J. *J. Chem. Phys.* **1992**, *96*, 6796–6806.
- (68) Woon, D. E.; Dunning, T. H., Jr. *J. Chem. Phys.* **1993**, *98*, 1358–1371.
- (69) Balabanov, N. B.; Peterson, K. A. *J. Chem. Phys.* **2005**, *123*, 64107.
- (70) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, N. J.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, Ö.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. *Gaussian 09*, Revision A.1.; Gaussian Inc.: Wallingford, CT, 2009.
- (71) Schäfer, A.; Horn, H.; Ahlrichs, R. *J. Chem. Phys.* **1992**, *97*, 2571–2577.
- (72) Schäfer, A.; Huber, C.; Ahlrichs, R. *J. Chem. Phys.* **1994**, *100*, 5829–5835.
- (73) Clopath, P.; von Zelewsky, A. *Helv. Chim. Acta* **1972**, *55*, 52–67.
- (74) Flory, M. A.; McLamarras, S. K.; Ziurys, L. M. *J. Chem. Phys.* **2006**, *125*, 194304.
- (75) Knight, J. B., Jr.; Mouchet, A.; Beaudry, W. T.; Duncan, M. *J. Magn. Reson.* **1978**, *32*, 383–390.
- (76) Belanzoni, P.; van Lenthe, E.; Baerends, E. J. *J. Chem. Phys.* **2001**, *114*, 4421–4433.
- (77) Malkin, I.; Malkina, O. L.; Malkin, V. G.; Kaupp, M. *Chem. Phys. Lett.* **2004**, *396*, 268–276.
- (78) Belanzoni, P.; Baerends, E. J.; van Asselt, S.; Langewen, P. B. *J. Phys. Chem.* **1995**, *99*, 13094–13102l.
- (79) Belanzoni, P.; Baerends, E. J.; Gribnau, M. *J. Phys. Chem. A* **1999**, *103*, 3732–3744.
- (80) DeVore, T. C.; Weltner, W., Jr. *J. Am. Chem. Soc.* **1977**, *99*, 4700–4703.
- (81) Hastie, J. W.; Hauge, R. H.; Margrave, J. L. *J. Chem. Phys.* **1969**, *51*, 2648–2656.

- (82) Carver, G.; Bendix, J.; Tregenna-Piggot, P. L. W. *Chem. Phys.* **2002**, *282*, 245–263.
- (83) Sygusch, J. *Acta Crystallogr., Sect. B* **1974**, *30*, 662–665.
- (84) Nugent, W. A.; Mayer, J. M. *Metal-Ligand Multiple Bonds: The Chemistry of Transition Metal Complexes Containing Oxo, Nitrido, Imido, Alkylidene, or Alkylidyne Ligands*; Wiley: New York, 1988.
- (85) Dodge, R. P.; Templeton, D. H. *J. Chem. Phys.* **1969**, *35*, 55–67.
- (86) Gahan, B.; Garner, C. D.; Hill, L. H.; Mabbs, F. E.; Hargave, K. D.; McPhail, A. T. *J. Chem. Soc., Dalton Trans.* **1977**, 1726–1729.
- (87) Campbell, R. F.; Freed, K. F. *J. Phys. Chem.* **1980**, *84*, 2668–2680.
- (88) Garner, C. D.; Hillier, I. H.; Mabbs, F. E.; Taylor, C.; Guest, M. F. *J. Chem. Soc., Dalton Trans.* **1976**, 2258.
- (89) Saladino, A. C.; Larsen, S. C. *J. Phys. Chem. A* **2003**, *107*, 1872–1878.
- (90) Kondo, M.; Minakoshi, S.; Iwata, K.; Shimizu, T.; Matsuzaka, H.; Kamigata, N.; Kitagawa, S. *Chem. Lett.* **1996**, 489–490.
- (91) Fairhurst, S. A.; Morton, J. R.; Preston, K. F. *Chem. Phys. Lett.* **1984**, *104*, 112–114.
- (92) Ferrante, R. F.; Wilkerson, J. L.; Graham, W. R. M.; Weltner, W., Jr. *J. Chem. Phys.* **1977**, *67*, 5904–5913.
- (93) Munzarová, M.; Kaupp, M. *J. Am. Chem. Soc.* **2000**, *122*, 11900–11913.
- (94) Bendix, J.; Meyer, K.; Weyhermüller, T.; Bill, E.; Metzler-Nolte, N.; Wieghardt, K. *Inorg. Chem.* **1998**, *37*, 1767–1775.
- (95) Davies, S. C.; Hughes, D. L.; Leigh, J. R.; Sanders, G. J.; de Souza, J. S. *J. Chem. Soc., Dalton Trans.* **1997**, 1981–1988.
- (96) Davies, G. R.; Jarvis, J. A. J.; Kilbourn, B. T.; Mais, R. H. B.; Owston, P. G. *J. Chem. Soc. A* **1970**, 1275–1283.
- (97) Feltham, R. D.; Crain, H. *Inorg. Chem. Acta* **1980**, *40*, 37–40.
- (98) Dethlefsen, J. W.; Hedegård, E. D.; Rimmer, R. D.; Ford, P. C.; Døssing, A. *Inorg. Chem.* **2009**, *48*, 231–238.
- (99) Rudin, M.; Schweiger, A.; Berchten, N.; Günthard, H. H. *Mol. Phys.* **1980**, *41*, 1317.
- (100) Rudin, M.; Jörin, E.; Schweiger, A.; Günthard, H. H. *Chem. Phys. Lett.* **1979**, *67*, 374–376.
- (101) Symons, M. C. R.; Bratt, S. W. *J. Chem. Soc., Dalton Trans.* **1979**, 1739–1743.
- (102) Cariati, F.; Morazzoni, F.; Busetto, C.; Del Piero, G.; Zazzetta, A. *J. Chem. Soc., Dalton Trans.* **1976**, 342–347.
- (103) Atanasov, M.; Baerends, E. J.; Baettig, P.; Bruyndonckx, R.; Daul, C.; Rauzy, C.; Zbiri, Z. *Chem. Phys. Lett.* **2004**, *399*, 433.
- (104) Zbiri, M. *Inorg. Chim. Acta* **2006**, *359*, 3865–3870.
- (105) Braden, D. A.; Tyler, D. R. *J. Am. Chem. Soc.* **1998**, *120*, 942–947.
- (106) Kobayashi, A.; Sasaki, Y. *Bull. Chem. Soc. Jpn.* **1977**, *50*, 2650–2656.
- (107) Maki, A. H.; Edelstein, N.; Davison, A.; Holm, R. H. *J. Am. Chem. Soc.* **1964**, *86*, 4580–4587.
- (108) Schmitt, R. D.; Holm, R. H. *J. Am. Chem. Soc.* **1968**, *90*, 2288–2292.
- (109) Huyett, J. E.; Choudhury, S. B.; Eichhorn, D. M.; Bryngelson, P. A.; Maroney, M. J.; Hoffman, B. M. *Inorg. Chem.* **1998**, *37*, 1361–1367.
- (110) Stein, M.; van Lenthe, E.; Baerends, E. J.; Lubitz, W. *J. Phys. Chem. A* **2001**, *105*, 416–425.
- (111) Stadler, C.; de Lacey, A. L.; Hernández, B.; M., F. V.; Conesa, J. C. *Inorg. Chem.* **2002**, *41*, 4417–4423.
- (112) Swink, L. N.; Atoji, M. *Acta Crystallogr.* **1960**, *13*, 639–643.
- (113) Kasai, P. H.; Jones, P. M. *J. Am. Chem. Soc.* **1985**, *107*, 813–818.
- (114) Huber, H.; Kündig, E. P.; Moskovits, M.; Ozin, G. A. *J. Am. Chem. Soc.* **1975**, *97*, 2097–2106.
- (115) Sastry, B. A.; Asadullah, S. M.; Ponticelli, P.; Massacei, M. *Spectrochim. Acta* **1979**, *35A*, 817–821.
- (116) Maxcy, K. R.; Turnbull, M. M. *Acta Crystallogr., Sect. C* **1999**, *C55*, 1986–1988.
- (117) Neese, F. *J. Phys. Chem. A* **2001**, *105*, 4290–4299.
- (118) Gauld, J. W.; Erikson, L. A.; Radom, L. *J. Chem. Phys.* **1997**, *101*, 1352–1359.

Essential on the Photophysics and Photochemistry of the Indole Chromophore by Using a Totally Unconstrained Theoretical Approach

Angelo Giussani,^{*,†} Manuela Merchán,[†] Daniel Roca-Sanjuán,^{*,‡} and Roland Lindh[‡]

[†]Instituto de Ciencia Molecular, Universitat de València, Apartado 22085, ES-46071 Valencia, Spain

[‡]Department of Chemistry – Ångström, Theoretical Chemistry Program, Uppsala University Box 518, 75120 Uppsala, Sweden

 Supporting Information

ABSTRACT: Indole is a chromophore present in many different molecules of biological interest, such as the essential amino acid tryptophan and the neurotransmitter serotonin. On the basis of CASPT2//CASSCF quantum chemical calculations, the photophysical properties of the system after UV irradiation have been studied through the exploration of the potential energy hypersurfaces of the singlet and triplet low-lying valence excited states. In contrast to previous studies, the present work has been carried out without imposing any restriction to the geometry of the molecule (C_1 symmetry) and by performing minimum energy path calculations, which is the only instrument able to provide the lowest-energy evolution of the system. Relevant findings to the photophysics of bare indole have been obtained, which compete with the currently accepted mechanism for the energy decay in the molecule. The results show the presence of a conical intersection (CI) between the initially populated $^1(L_a \pi\pi^*)$ and the $^1(L_b \pi\pi^*)$ state, easily accessible through a barrierless pathway from the Franck–Condon region. At this CI region, part of the population is switched from the bright $^1(L_a \pi\pi^*)$ state to the $^1(L_b \pi\pi^*)$ state, and the system evolves toward a minimum structure from which the expected fluorescence takes place. The reported low values of the fluorescence quantum yield are explained by means of a new nonradiative mechanism specific for the $^1(L_b \pi\pi^*)$ state, in which the presence of an ethene-like CI between the $(L_b \pi\pi^*)$ and ground states is the main feature.

INTRODUCTION

Due to their omnipresence in biological systems, the research on proteins has always been a field of enormous importance in almost every branch of bioscience. Regarding their photophysical properties, the main features of the near and far ultraviolet spectra of proteins have been reported in the regions between 220 and 190 nm (5.64–6.53 eV) and around 280 nm (4.43 eV).^{1,2} This behavior is explained in terms of their constituents and in particular by the presence of aromatic amino acids [phenylalanine (Phe), tyrosine (Tyr), tryptophan (Trp), histidine (His)] whose photophysics determine the global response of proteins to UV radiation. Within the cited aromatic amino acids, the study of tryptophan has attracted special attention since it undoubtedly constitutes the most important emissive source in polypeptides, and its photophysical response is highly sensitive to the local microenvironment. The study of tryptophan is then relevant both to explain the global photophysics of many different proteins and to evaluate the potential of this amino acid as a useful probe of local environment and dynamics in proteins.³ The chromophore group of tryptophan is indole (see Figure 1), which is also present in many different molecules of biological interest, like the neurotransmitter serotonin. A complete and correct description of the photophysical behavior of indole is consequently not only a matter of fundamental knowledge, but it constitutes the basis toward the understanding of the UV-induced processes taking place in more complicated related compounds.

Many experimental^{4–20} and theoretical^{21–27} studies have been performed on bare indole. The analysis of the gas-phase absorption spectrum of the molecule in the low-energy UV range has shown the contribution of two electronic transitions, one responsible for the

sharp peak at 4.37 eV^{4–8} and the other related to a broad continuum peaking around 4.77 eV.^{7,9} These transitions have been assigned to two low-lying $\pi\pi^*$ valence singlet excited states which have been labeled, according to the Platt's nomenclature,²⁸ as $^1(L_b \pi\pi^*)$ and $^1(L_a \pi\pi^*)$. The $^1(L_a \pi\pi^*)$ state is characterized by a higher dipole moment, 5.4 D,¹¹ in comparison with the values for $^1(L_b \pi\pi^*)$ and the ground state, equal to 2.3¹² and 2.09 D,¹³ respectively. Theoretical calculations were able to properly describe the absorption maxima measured in the experiments. Serrano-Andrés and Roos,²² by using the complete active space self-consistent field/multiconfigurational second-order perturbation theory (CASSCF/CASPT2) methodology, reported the vertical excitation energies of the $^1(L_b \pi\pi^*)$ and $^1(L_a \pi\pi^*)$ states at 4.43 and 4.73 eV, respectively, in agreement with the experimental data. Regarding band origins, the value related to the $^1(L_b \pi\pi^*)$ state was identified by means of the analysis of the fluorescence excitation and emission spectra of the system and is considered to be coincident to the reported absorption maximum at 4.37 eV.⁷ The $^1(L_a \pi\pi^*)$ band origin was more problematic, since its location was for a long time hampered by the vibronic structure of the $^1(L_b \pi\pi^*)$ state. Finally, the value was measured at 4.54 eV by means of site-selected fluorescence excitation in solid Ar matrices.¹⁰ Such data have been also corroborated by the theoretical works of Serrano-Andrés, Roos, and Borin.^{22,25}

Both theoretical and experimental studies on indole indicate that most of the molecules reach the $^1(L_a \pi\pi^*)$ state after UV

Received: September 15, 2011

Published: November 07, 2011

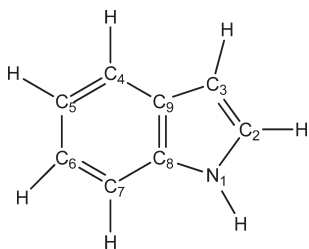


Figure 1. Indole structure and atom labeling.

irradiation. The oscillator strength (f) associated to the corresponding transition has been found to be larger than the value obtained for the $^1(L_b, \pi\pi^*)$ state.^{22,24} Subsequently, the population of the initially bright state is transferred to the $^1(L_b, \pi\pi^*)$ state, as previously shown by Brand et al.,^{20,21} which is expected to be the responsible for the fluorescence spectra of indole in the gas phase.²² An interesting photophysical property of the system observed in the experiments is the dramatic decrease of the fluorescence quantum yield with the increase of the excitation energy.¹⁴ This points out the presence of an efficient nonradiative decay path only accessible when the molecule is irradiated with an extra energy. In order to explain such behavior, Sobolewski, Domcke, and co-workers^{23,26,27} proposed a nonradiative relaxation mechanism based on the topology of the potential energy hypersurfaces (PEHs) of a dark $\pi\sigma^*$ electronic state as a function of the proton-detachment coordinate of the NH group, which provides a plausible radiationless decay path whose importance has been also highlighted experimentally.^{15,16}

Such global scenario of the radiative and nonradiative properties of indole has been accepted by many theoretical and experimental groups during the last decades, although a totally unconstrained theoretical research on the evolution of the initially populated $^1(L_a, \pi\pi^*)$ state has not been carried out. Most of the cited works were in fact performed by constraining the molecule to planarity, i.e., imposing C_s symmetry, and no minimum energy paths (MEPs)^{29,30} have been accounted for. In the studies carried out by Serrano-Andrés, Roos, and Borin,^{22,24,25} the geometry for the emissive species was obtained by the standard optimization procedure, a strategy not able to show the presence or not any barrier between the Franck–Condon (FC) region and the excited-state minimum structures. On the other hand, Sobolewski, Domcke, and co-workers^{23,26,27} explored the PEHs profile only along the proton-detachment coordinate of the NH group, through the optimization of different structures each with a fixed value of the NH bond distance.

In the present contribution, by using the current state-of-the-art computational strategies, we have obtained the unconstrained evolution of the system, allowing to relax all possible degrees of freedom of the molecule. The MEP technique with mass-weighted coordinates has been employed. This procedure is in fact the only computational tool able to describe the adiabatic evolution of a state, providing (if present) a steepest descendent path which consequently will constitute the most favorable decay experimented by the system. Our results clearly identify as primary process upon the population of the brightest excited state [the $^1(L_a, \pi\pi^*)$ state] a switch to the lowest excited state $^1(L_b, \pi\pi^*)$ driven by the presence of a conical intersection, reported recently by Brand et al.^{20,21} In addition, the analysis performed without imposing any symmetry constraint to the system allows to find and characterize a nonradiative decay

process particular to the $^1(L_b, \pi\pi^*)$ state, which is mediated by an nonplanar ethene-like CI^{31,32,39} between $^1(L_b, \pi\pi^*)$ and the ground state. This radiationless decay path rationalizes the experimental observations, complementing the current established mechanism for energy-decay in the indole molecule, and contributes therefore to the elucidation of the global response of bare indole to UV radiation.

The results are presented in five sections. First, the low-lying singlet and triplet excited states involved in the photophysics of indole are analyzed. In the second section, evolution of the system after excitation to the brightest low-lying excited state without any excess energy is elucidated by the description of the so-called main decay path. Next, two additional decay paths are described. In section four, a new nonradiative decay mechanism specific for the $^1(L_b, \pi\pi^*)$ state is considered in detail. Finally, the photophysics of the two low-lying triplet valence excited states is briefly discussed.

METHODOLOGY

The present study has been performed by using the well-tested CASPT2//CASSCF methodology^{33–36} as implemented in the MOLCAS 7.4 software.³⁷ Optimized structures and minimum energy paths have been then calculated at the multi-configurational CASSCF level, and at the geometries so obtained, the dynamic correlation effects have been taken into account in the energies by performing second-order multi-configurational CASPT2 calculations. All computations have been performed by imposing no restrictions to the symmetry of the molecule (C_1 symmetry); out-of-plane geometry distortions were therefore allowed. Two basis sets of atomic natural orbital (ANO) of S- and L-type contracted to C,N [4s,3p,1d]/H[2s1p], have been employed resulting in similar conclusions. The results reported here correspond to the higher level of theory, that is, employing the latter basis set. The whole π system of the molecule has been considered in the active space. Thus, it comprises six π orbitals of the benzene ring plus two π orbitals of the pyrrole ring and the π orbital of the nitrogen atom, together with the corresponding 10 electrons [CASSCF(10,9)]. As the CASPT2//CASSCF methodology is strongly determined not only by the size of the active space but also by the shape of the orbitals employed, for the sake of clarity, a picture of the active orbitals is reported in Figure S1, Supporting Information. Within the CASPT2 calculations, an imaginary level-shift correction of 0.2 au has been used in order to avoid the presence of intruder states. The CASPT2 standard zeroth-order Hamiltonian has been used as originally implemented.³⁴ The core orbitals have been frozen in the CASPT2 calculations. Such CASPT2 approach has been validated during the last decades in many different studies on organic molecules providing a correct description and interpretation of the photophysical experimental data.^{38,39} As highlighted above, research of the evolution of the valence excited states has been performed by means of MEP calculations.^{29,30} Mass-weighted coordinates have been used. This technique provides (if present) a steepest descendent path, in which each step is built by the minimization of the energy on a hyperspherical cross section of the PEH centered on the initial geometry within a predefined radius. The importance and reliability of results obtained using such a computational tool have been proven in many different studies which confirm MEP calculations as a valuable procedure for the description of the photophysics and photochemistry of a molecule.^{40,41} CIs not obtained along MEPs

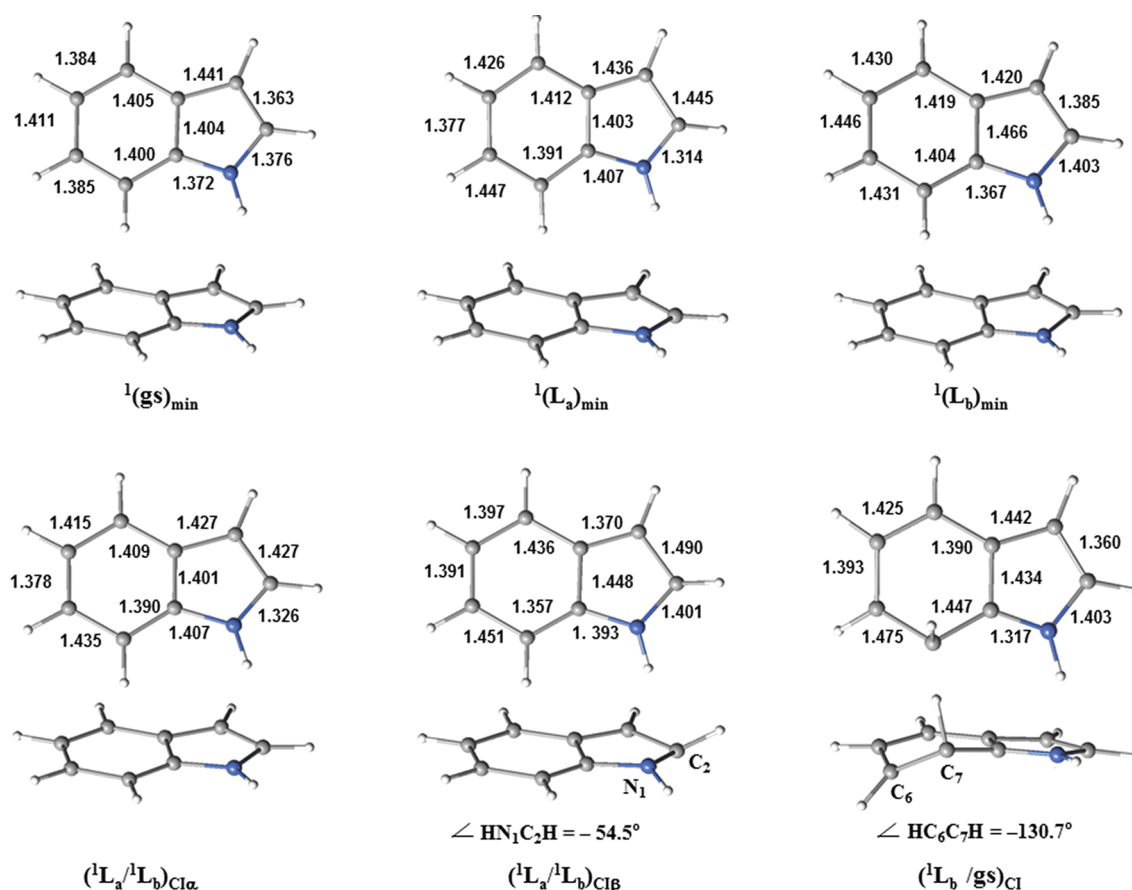


Figure 2. Frontal and side views of the CASSCF optimized structures of indole. CC and CN bond lengths (in Å) are also shown.

calculations have been computed by using the restricted Lagrange multipliers technique, as included in the MOLCAS 7.4 package in which the lowest-energy point is obtained under the restriction of degeneracy between the two considered states.³⁷

RESULTS AND DISCUSSION

FC Geometry: Singlet and Triplet Valence Excited States. Optimization of the ground state of indole with a totally unconstrained approach leads to a planar geometry, labeled $^1(\text{gs})_{\text{min}}$. The frontal and side views of this structure are depicted in Figure 2, and the most significant bond lengths and angles are compiled in Table 1. All these structural parameters are in agreement with previous results obtained by Serrano-Andrés and Roos.²² At the geometry of this minimum, $^1(\text{gs})_{\text{min}}$, a vertical CASPT2//CASSCF calculation of the low-lying six singlet excited states has been undertaken in order to study the absorption properties of the molecule in the FC region. For the sake of completeness, the lowest triplet excited states have been also computed. Only valence excited states have been considered since the present study is focused on the photophysics of indole, which is mainly determined by the valence electronic structure of the molecule. Analysis of nonpure valence excited states has been provided elsewhere.^{22,23,26,27} For instance, Serrano-Andrés and Roos studied the nature of several low-lying excited states and the corresponding excitation energies by using extended basis sets, including Rydberg basis functions.²² The two lowest excited states, which will be the ones studied in the present work, have been determined to be valence states. The analysis

of the orbitals involved in the excitations related to these low-lying valence excited states has been carried out here to establish their nature. According to the Platt's nomenclature,²⁸ the valence states of interest can be described in terms of the natural orbitals (NOs) topologically equivalent to the highest occupied molecular orbital (HOMO) and lowest unoccupied molecular orbital (LUMO). Thus, the CASSCF wave function basically described by the minus linear combination of the HOMO (H) \rightarrow LUMO (L) + 1 (25%) and H - 1 \rightarrow L (47%) configurations is identified as the $^1(\text{L}_b \pi\pi^*)$ state, while the CASSCF wave function mainly composed by the H \rightarrow L (56%) one-electron promotion is recognized as the $^1(\text{L}_a \pi\pi^*)$ state. The computed CASPT2 energies show that the $^1(\text{L}_b \pi\pi^*)$ and $^1(\text{L}_a \pi\pi^*)$ states are the lowest $\pi\pi^*$ valence singlet excited states, placed vertically (E_{VA}) at 4.36 and 4.79 eV, respectively (see Table 2). The results are in agreement with the available experimental data.^{1,2,4,5,7,9} Table 3 compiles in addition the computed oscillator strengths (f) and dipole moments (μ) for the $^1(\text{L}_b \pi\pi^*)$ and $^1(\text{L}_a \pi\pi^*)$ states. According to the results, the latter is predicted to be the bright state upon UV irradiation. The calculated dipole moment, μ , at the FC region is also consistent with the experimental findings.^{11–13} The $^1(\text{L}_b \pi\pi^*)$ state has a low μ (1.55 D), similar to the value obtained for the ground state (1.81 D), while $^1(\text{L}_a \pi\pi^*)$ is characterized by a high μ (6.07 D), which is more than three times larger with respect to the former.

Rydberg states are not described here, and for this reason, the basis set employed in our calculations does not include the

Table 1. Calculated and Experimental Bond Lengths and Angles for the Optimized Ground-State Geometry ${}^1(\text{gs})_{\text{min}}$, the Conical Intersection $({}^1\text{L}_a/{}^1\text{L}_b)_{\text{CI}\alpha}$, and the Equilibrium Geometry ${}^1(\text{L}_b)_{\text{min}}$ of the Indole Molecule

bonds ^a	${}^1(\text{gs})_{\text{min}}^b$	$({}^1\text{L}_a/{}^1\text{L}_b)_{\text{CI}\alpha}$	${}^1(\text{L}_b)_{\text{min}}^b$	exptl ^c	angles ^a	${}^1(\text{gs})_{\text{min}}^b$	$({}^1\text{L}_a/{}^1\text{L}_b)_{\text{CI}\alpha}$	${}^1(\text{L}_b)_{\text{min}}^b$	exptl ^c
N ₁ C ₂	1.376	1.326	1.403	1.377	N ₁ C ₂ C ₃	109.6	107.7	108.4	111.5
C ₂ C ₃	1.363	1.427	1.385	1.344	C ₂ C ₃ C ₉	106.7	106.8	107.8	105.5
C ₃ C ₉	1.441	1.427	1.420	1.451	C ₃ C ₉ C ₄	133.9	132.9	133.5	132.2
C ₉ C ₄	1.405	1.409	1.419	1.412	C ₉ C ₄ C ₅	118.9	116.9	118.0	114.6
C ₄ C ₅	1.384	1.415	1.430	1.397	C ₄ C ₅ C ₆	120.9	122.3	121.7	124.8
C ₅ C ₆	1.411	1.378	1.446	1.386	C ₅ C ₆ C ₇	121.2	121.5	120.8	119.7
C ₆ C ₇	1.385	1.435	1.431	1.399	C ₆ C ₇ C ₈	117.5	115.0	117.1	116.4
C ₇ C ₈	1.400	1.390	1.404	1.400	C ₈ N ₁ H	125.7	124.1	125.0	
C ₈ N ₁	1.372	1.407	1.367						
C ₈ C ₉	1.404	1.401	1.466						
N ₁ H	0.990	0.997	0.989						

^a Bonds are in Å and angles in degrees (°). ^b ANO-L type basis set C,N [4s,3p,1d]/H[2s1p], CASPT2//CASSCF(10,9) calculations. ^c Taken from the crystal X-ray structure for the ground state of tryptophan (ref 42).

Table 2. Calculated and Experimental Energy (in eV) for the Low-Lying Excited Valence States of the Isolated Indole Molecule

state	theory			experiments ^c		
	E_{VA}^a	T_e^a	E_{VE}^a	A_{max}^b	T_0^b	R. emi. ^b
${}^1(\text{L}_b \pi\pi^*)$	4.36	4.11	4.01	4.37	4.37	4.12 ^e /4.36
${}^1(\text{L}_a \pi\pi^*)$	4.79	4.40	4.07	4.77	4.54	
${}^3(\text{L}_a \pi\pi^*)$	3.42	3.01	2.68	3.3 ^d	3.07 ^f	2.87 ^f
${}^3(\text{L}_b \pi\pi^*)$	4.05	3.83	3.66			

^a ANO-L type basis set C,N [4s,3p,1d]/H[2s1p], CASPT2//CASSCF-(10,9) vertical excitation energy (E_{VA}), electronic band origin (T_e), and vertical emission energy (E_{VE}). ^b Measured absorption band maximum (A_{max}), band origin (T_0), and relaxed emission from the excited-state relaxed geometry (R. emi.). ^c Experiments from ref 1, 2, 4, 5, 7, and 9. ^d Estimated excitation energy from the ground state minimum (see ref 22). ^e Experiments from ref 17. ^f Experimental spectra from ref 18 and 19.

required Rydberg-type functions to treat properly such states. Some calibration computations have been performed by locating at the center of mass Rydberg basis functions, which have been built specifically for the indole molecule (see Table S1, Supporting Information). The results obtained ensure an accurate description of the ${}^1(\text{L}_b \pi\pi^*)$ and ${}^1(\text{L}_a \pi\pi^*)$ states by means of the basis set without diffuse functions. In addition, the work carried out by Serrano-Andrés and Roos²² on the theoretical study of the absorption spectrum of indole proved that the lowest Rydberg state appears vertically at 4.85 eV with respect to the ground-state energy, and our calculations show that only the ${}^1(\text{L}_b \pi\pi^*)$ and ${}^1(\text{L}_a \pi\pi^*)$ singlet excited states have a E_{VA} lower than 5.00 eV. Consequently, only the evolution of these two low-lying singlet valence excited states is analyzed in the next sections.

Regarding triplet excited states, the five low-lying computed roots present E_{VA} values lower than 5.00 eV, and therefore these states have been taken into account in the photochemical description of the system. As for the singlet excitations, the nature of the triplet states has been identified by inspection of the respective CASSCF wave functions, and each of them has been labeled by using Platt's nomenclature.²⁸ The overall results for the triplet states, together with those related to ${}^1(\text{L}_b \pi\pi^*)$ and ${}^1(\text{L}_a \pi\pi^*)$, are summarized in Table 3. In contrast to the singlet

Table 3. Calculated Vertical Excitation Energies at the FC Geometry (E_{VA} , eV) for the Lowest Valence Singlet and Spin Forbidden Triplet Excited States^a

root	state	E_{VA}	f	μ
S ₀	gs			1.81
T ₁	${}^3(\text{L}_a \pi\pi^*)$	3.42		1.41
T ₂	${}^3(\text{L}_b \pi\pi^*)$	4.05		1.76
S ₁	${}^1(\text{L}_b \pi\pi^*)$	4.36	0.018	1.55
T ₃	${}^3(\text{B}_a \pi\pi^*)$	4.50		1.83
T ₄	${}^3(\pi\pi^*)$	4.74		1.46
T ₅	${}^3(\text{B}_b \pi\pi^*)$	4.76		5.08
S ₂	${}^1(\text{L}_a \pi\pi^*)$	4.79	0.078	6.07

^a The computed dipole moments (μ , D) and the oscillator strengths (f) for the singlet–singlet transitions are also included.

states, the ${}^3(\text{L}_a \pi\pi^*)$ has a lower E_{VA} than ${}^3(\text{L}_b \pi\pi^*)$, and both of them have dipole moments similar to that of the value obtained for the ground state.

Main Decay Path. From the analysis of the oscillator strength, it is possible to assert that most of the molecules reach the ${}^1(\text{L}_a \pi\pi^*)$ state after UV irradiation (the calculated f of the ${}^1(\text{L}_b \pi\pi^*)$ and ${}^1(\text{L}_a \pi\pi^*)$ states are 0.018 and 0.078, respectively). As stated above, from the FC geometry evolution of the initially populated ${}^1(\text{L}_a \pi\pi^*)$ state has been characterized by means of the MEP approach with no spatial symmetry constrains. The computational result is a barrierless path leading directly from the FC region to a CI involving the ${}^1(\text{L}_b \pi\pi^*)$ and ${}^1(\text{L}_a \pi\pi^*)$ states, placed at 4.42 eV with respect to the ground-state minimum, denoted hereafter as $({}^1\text{L}_a/{}^1\text{L}_b)_{\text{CI}\alpha}$ (see Figure 3). Such degenerate region is a converged point on the obtained MEP, and it is not the lowest-energy point of the crossing seam but the first funnel that can be reached by the excited molecule, and consequently, the most photophysical relevant crossing point.⁴³ This CI has been characterized recently by Brand et al.^{20,21} with an upper bound of around 4.6 eV at the DFT/MRCI level. An efficient internal conversion (IC) process mediated by this CI consequently causes the energy transfer from ${}^1(\text{L}_a \pi\pi^*)$ to ${}^1(\text{L}_b \pi\pi^*)$. Actually almost all the MEP points are characterized by a quite pronounced energy degeneracy between the ${}^1(\text{L}_b \pi\pi^*)$ and ${}^1(\text{L}_a \pi\pi^*)$ states, which suggests that along the decay pathway

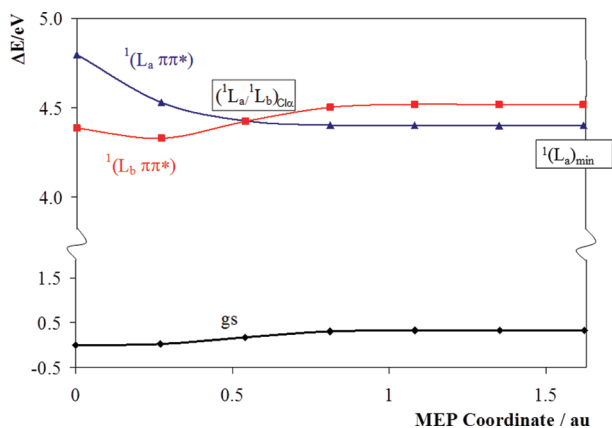


Figure 3. Evolution of the ground and two lowest singlet excited states for indole from the FC geometry along the ${}^1(L_a \pi\pi^*)$ MEP computed at the CASPT2//CASSCF(10,9)/ANO-L C,N [4s,3p,1d]/H[2s1p] level. The corresponding CASSCF MEP can be found in Figure S2, Supporting Information.

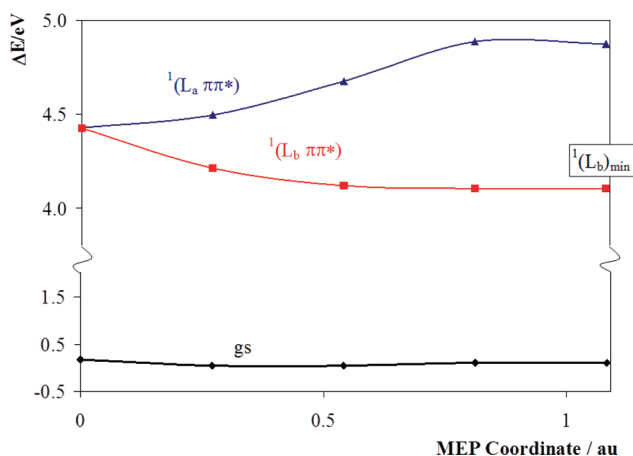


Figure 4. Evolution of the ground and two lowest singlet excited states for indole from the $({}^1L_a/{}^1L_b)_{CI\alpha}$ conical intersection along the ${}^1(L_b \pi\pi^*)$ MEP computed at the CASPT2//CASSCF(10,9)/ANO-L C,N [4s,3p,1d]/H[2s1p] level. The corresponding CASSCF MEP can be found in Figure S3, Supporting Information.

from the FC region, the probability of populating the ${}^1(L_b \pi\pi^*)$ state is high. The equilibrium structure itself of the ${}^1(L_a \pi\pi^*)$ state, obtained as the last point of the MEP calculation, hereafter ${}^1(L_a)_{min}$, is placed at only 0.12 eV below ${}^1(L_b \pi\pi^*)$ state and is almost degenerated with the $({}^1L_a/{}^1L_b)_{CI\alpha}$ point which mediates the ultrafast relaxation to the ${}^1(L_b \pi\pi^*)$ state. According to these findings, the ${}^1(L_a)_{min}$ structure, with a computed vertical emission (E_{VE}) of 4.07 eV (see Table 2), is not expected to contribute markedly to the fluorescence phenomenon in the indole molecule. This is in contrast with the initial evidence of dual fluorescence in both polar and nonpolar solvents, based on measurements of the polarization of the fluorescence bands,⁴⁴ and supports further experimental findings suggesting that, while two different states seem to contribute in the lowest-energy absorption band, only one of them is responsible for the emission.^{2,45}

Figure 4 displays the results obtained for the MEP from $({}^1L_a/{}^1L_b)_{CI\alpha}$. As can be seen in the picture, the ${}^1(L_b \pi\pi^*)$ state

evolves through a steepest descendent path to the equilibrium geometry, ${}^1(L_b)_{min}$. Similar results are obtained by carrying out a MEP from the ${}^1(L_a)_{min}$ structure. At the ${}^1(L_b)_{min}$ geometry, the singlet excited state ${}^1(L_b \pi\pi^*)$ is the lowest on the S_1 hypersurface, and since the system cannot further evolve in a radiationless manner, we conclude that it is from this region that the emission takes place. For the S_1 state, the computed E_{VA} , the electronic band origin (T_e), and E_{VE} are 4.36, 4.11, and 4.01 eV, respectively (see Table 2). The present results agree within 0.26 eV with those obtained by Serrano-Andrés and Roos,²² being particularly closer to the experimental data reported by Borisevich et al.¹⁷ than to other initial observations.⁷ In general all the theoretical and experimental studies result in lower E_{VA} , T_e , and E_{VE} values for ${}^1(L_b \pi\pi^*)$ with respect to the ${}^1(L_a \pi\pi^*)$ state.

The present optimized ${}^1(L_a)_{min}$, $({}^1L_a/{}^1L_b)_{CI\alpha}$ and ${}^1(L_b)_{min}$ structures are almost planar (see Figure 2 and Table 1). In both ${}^1(L_a)_{min}$ and $({}^1L_a/{}^1L_b)_{CI\alpha}$, the relative positions of the single and double bonds are interchanged with respect to ${}^1(gs)_{min}$. Meanwhile the ${}^1(L_b)_{min}$ geometry reflects electron delocalization taking place within the benzene ring.

Additional Decay Paths. In the previous section, the main decay path on the photophysics of indole has been characterized, i.e., the primary process that follows the population of the brightest excited state upon light irradiation. To improve the description, two other relevant deactivation paths have been studied and are next analyzed.

The first one is based on the fact that the ${}^1(L_b \pi\pi^*)$ singlet state can also be initially populated by UV absorption, since the computed f is not negligible. In order to study such possibility, a MEP on the ${}^1(L_b \pi\pi^*)$ singlet state has been computed starting from the FC structure (see Figures 5a). The final point of the MEP corresponds to the same region previously found along the main decay path: the ${}^1(L_b)_{min}$ structure. This new evidence confirms the relevant participation of the equilibrium structure of the ${}^1(L_b \pi\pi^*)$ state, ${}^1(L_b)_{min}$, in the fluorescence phenomenon of indole.

The second possible deactivation path studied involves an alternative CI between the ${}^1(L_a \pi\pi^*)$ and ${}^1(L_b \pi\pi^*)$ states, which will be denoted as $({}^1L_a/{}^1L_b)_{CI\beta}$. This CI region might be important in the photophysics of indole since its energy with respect to the ground-state minimum is equal to 4.54 eV, which is only 0.12 eV above the former CI, $({}^1L_a/{}^1L_b)_{CI\alpha}$. The new structure, $({}^1L_a/{}^1L_b)_{CI\beta}$, is therefore placed in a region of the PEH energetically close to the main decay path that the system is expected to undertake upon absorption of UV light. As shown in Figure 2, $({}^1L_a/{}^1L_b)_{CI\beta}$, in contrast to $({}^1L_a/{}^1L_b)_{CI\alpha}$ is characterized by a pronounced out-of-plane distortion mainly localized on the HN_1 and C_2H groups of atoms; the dihedral angle related to these atoms is equal to -54.5° .

In order to establish the accessibility of this region, which would give us a hint on the relevance of the CI for the photophysics of the system, a linear interpolation of internal coordinates (LIIC) calculation has been performed on both the ${}^1(L_a \pi\pi^*)$ and the ${}^1(L_b \pi\pi^*)$ states between the FC and $({}^1L_a/{}^1L_b)_{CI\beta}$ structures (see Figure 5b). Such a type of calculation is able to provide a connected path between two regions of the PEH, along with the corresponding (if present) barrier, which constitutes an upper bound to the energy required to reach the final geometry from the starting one. The so-computed energy barriers along the ${}^1(L_a \pi\pi^*)$ and ${}^1(L_b \pi\pi^*)$ manifolds from the FC region to the $({}^1L_a/{}^1L_b)_{CI\beta}$ geometry are 0.43 and 0.26 eV, respectively. These values show that the $({}^1L_a/{}^1L_b)_{CI\beta}$

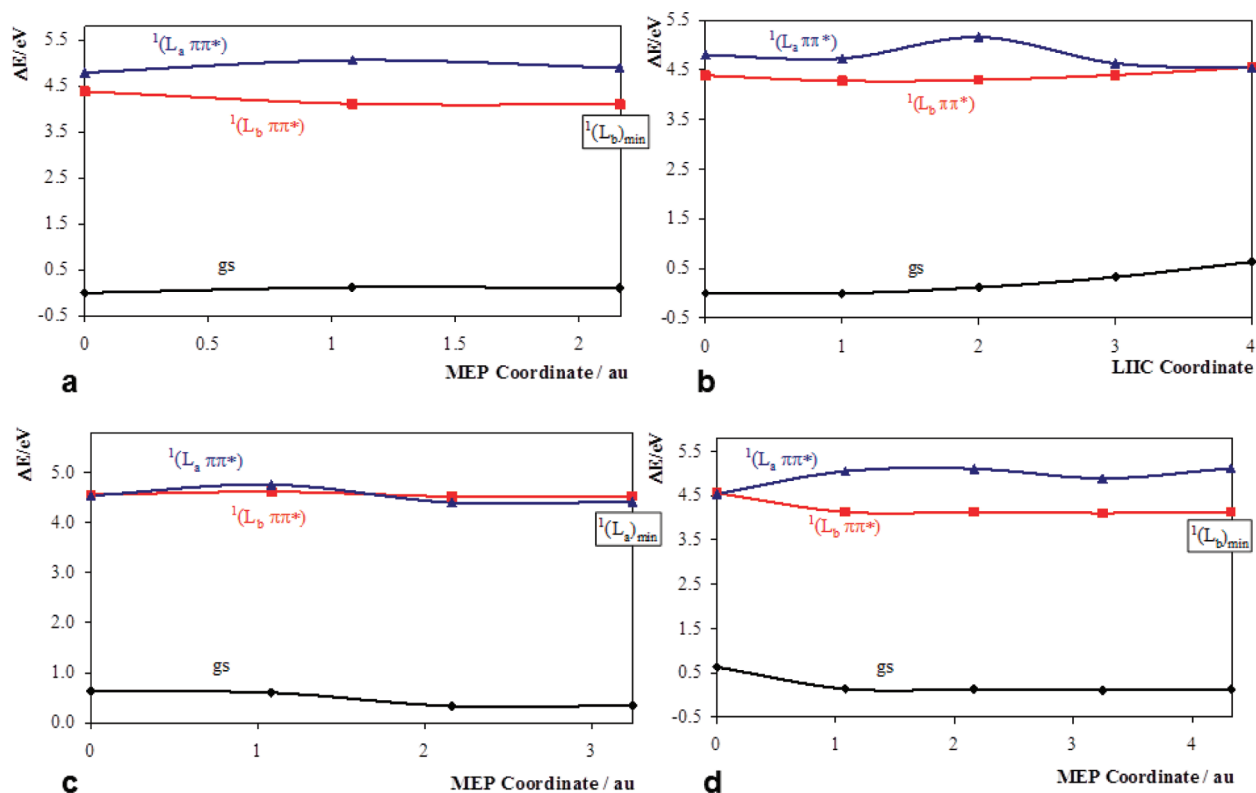


Figure 5. Energy profiles of the ground and two lowest singlet excited states for indole along the additional decay paths computed at the CASPT2//CASSCF(10,9)/ANO-L C,N [4s,3p,1d]/H[2s1p] level: (a) $^1(L_b \pi\pi^*)$ MEP from the FC geometry, $^1(gs)_{\min}$; (b) LIIC between $^1(gs)_{\min}$ and $(^1L_a/^1L_b)_{CI\beta}$; (c) $^1(L_a \pi\pi^*)$ MEP from $(^1L_a/^1L_b)_{CI\beta}$; and (d) $^1(L_b \pi\pi^*)$ MEP from $(^1L_a/^1L_b)_{CI\beta}$. The corresponding CASSCF MEPs can be found in Figures S4–S6, Supporting Information, respectively.

region might be accessible after UV-irradiation at least from the $^1(L_b \pi\pi^*)$ state. An energy barrier of 0.1–0.2 eV is in fact normally considered a small barrier, although only dynamics studies of the system can actually provide such conclusion.

According to the suggested plausible accessibility of the $(^1L_a/^1L_b)_{CI\beta}$ region upon UV irradiation, the evolution of the system from this CI has been also studied to give a complete description of the photophysics of bare indole. A MEP calculation shows that the $^1(L_a \pi\pi^*)$ state will evolve from the $(^1L_a/^1L_b)_{CI\beta}$ region to the previously characterized $^1(L_a)_{\min}$ structure (see Figure 5c). Although the path on the $^1(L_a \pi\pi^*)$ PEH connecting the structures has been obtained through the use of the MEP technique, the correction introduced in energies by means of the CASPT2 computations results in a small barrier of 0.22 eV. This result further points out the importance of taking into account the correlation effects associated with the instantaneous short-range electron–electron interaction in order to give a correct description of excited-state PEHs. Due to the low value of the barrier obtained at the CASPT2 level, we conclude that the $^1(L_a \pi\pi^*)$ state will decay to the $^1(L_a)_{\min}$ structure, from which, as previously stated, the system might evolve to the $^1(L_b)_{\min}$ equilibrium geometry. For the sake of completeness, from the $(^1L_a/^1L_b)_{CI\beta}$ point the behavior of the $^1(L_b \pi\pi^*)$ state has also been studied by means of computing the corresponding MEP (Figure 5d). The evolution of the $^1(L_b \pi\pi^*)$ state from the $(^1L_a/^1L_b)_{CI\beta}$ region leads to an equilibrium geometry which belongs again to the region on the PEH of S_1 previously identified as responsible for the fluorescence produced by the system: the $^1(L_b)_{\min}$ emissive structure.

In summary, all the secondary photophysical events analyzed, derived from the population of the $^1(L_b)$ state at the FC region or from the access of the $(^1L_a/^1L_b)_{CI\beta}$ structure after light absorption, drive the system toward the equilibrium region on the S_1 PEH, $^1(L_b)_{\min}$, which is assigned the fluorescence feature of bare indole.

Nonradiative Decay Process toward the Ground State. We have shown that the photophysics of indole is characterized by different decay paths leading to the $^1(L_b)_{\min}$ equilibrium structure, from which the molecule is predicted to emit light. Nevertheless, the intensity of the fluorescence band has been observed in several experiments to be sensitive to the excitation energy and the experimental conditions. In particular, the fluorescence quantum yield of bare indole drastically decreases as the excitation energy provided to the system increases.¹⁴ The loss of fluorescence emission in bare indole was previously explained by Sobolewski, Domcke, and co-workers^{23,26,27} who have shown that the presence of a dark $\pi\sigma^*$ state could be the responsibility of a nonradiative decay mechanism. According to their results, in order to activate such a radiationless process, the initially populated $^1(L_a \pi\pi^*)$ state has to reach a CI with the $\pi\sigma^*$ state which is placed above 5.00 eV with respect to the ground-state minimum. The authors then concluded that such barrier could explain the decrease of the fluorescence quantum yield of the system as the excitation energy is increased and predicted a threshold for this photophysical process to be located at about 5 eV for excitation from the minimum of the ground state. This mechanism is consistent with the increase of the nonradiative decay rate measured for indole with respect to

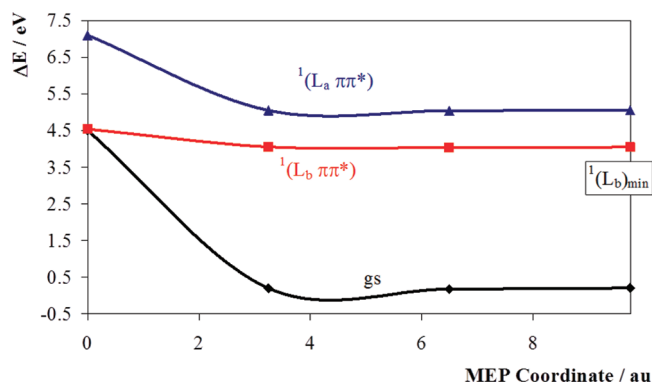


Figure 6. Evolution of the ground and two lowest singlet excited states for indole from the $(^1L_b/gs)_{CI}$ conical intersection along the $(^1L_b \pi\pi^*)$ MEP computed at the CASPT2//CASSCF(10,9)/ANO-L C,N [4s,3p,1d]/H[2s1p] level. The corresponding CASSCF MEP can be found in Figure S7, Supporting Information.

N-methylindole.¹⁴ However, it does not account for other experimental observations. The emission quantum yield of indole after excitation to the $(^1L_a \pi\pi^*)$ state at 257.2 nm (4.81 eV) is equal to 0.103,¹⁴ which is too low in comparison to the band intensity observed when excitation energies corresponding to the $(^1L_b \pi\pi^*)$ band origin are employed or to the present computational results.⁷ In fact, Ilich concluded in the experimental work on the low-lying singlet states of indole that the system has remarkable emission features only for energies around the $(^1L_b \pi\pi^*)$ band origin (measured at 4.37 eV), whereas a 22-fold decrease is observed when the molecule is provided with an extra energy of 4200 cm^{-1} (0.52 eV).⁷ Such experimental observations seem to point out the presence of a nonradiative decay path accessible for excitation energies lower than 5.00 eV and related to the $(^1L_b \pi\pi^*)$ state, apart from the radiationless route characterized by Sobolewski, Domcke, and co-workers.^{23,26,27}

Consistent with these experimental findings, we have found a CI between $(^1L_b \pi\pi^*)$ and the ground state, denoted hereafter as $(^1L_b/gs)_{CI}$. This degenerate region is placed at 4.50 eV with respect to $(^1gs)_{min}$ and is characterized by a strong out-of-plane distortion, mainly localized on the C₆H and C₇H group of atoms (see Figure 2). Such a crossing point has some geometrical similarities with the so-called triangular/kink S₁/S₀ CI reported for polyenes and aromatic systems,^{46,47} but it does not have the same electronics characteristics (three weakly coupled π -electrons on the out-of-plane distortion and a fourth electron localized along the remaining part of the benzene ring). It can be in fact better classified as a typical ethene-like CI,^{31,32,39} in which a twist of the double bond between the C₆ and C₇ atoms leads to an out-of-plane deformation with a HC₆C₇H dihedral angle of -127.1° and a pyramidalization of the C₇ atom. In order to determine the relevance of this CI on the photophysics of indole, a MEP on the S₁ state has been computed from the $(^1L_b/gs)_{CI}$ region (see Figure 6). This calculation shows the presence of a barrierless path leading to the $(^1L_b)_{min}$ region, which proves that $(^1L_b/gs)_{CI}$ can be reached from the equilibrium geometry of the $(^1L_b \pi\pi^*)$ state by overcoming a barrier of 0.39 eV. Such processes consequently constitute a nonradiative decay path specific for the $(^1L_b \pi\pi^*)$ state, whose accessibility depends on the energy provided to the system and can be activated at excitation energies lower than 5.00 eV.

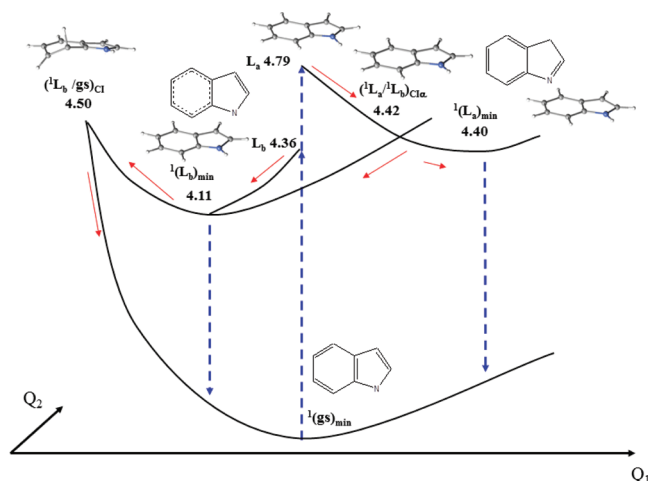


Figure 7. Scheme of the photophysics of indole along the singlet manifold. Energies (in eV) relative to the ground-state minimum structure, $(^1gs)_{min}$, are given. The Q_1 reaction coordinate is mainly related to the simple/double-bond rearrangements keeping the planarity of the molecule, meanwhile Q_2 is associated to out-of-plane distortions.

Figure 7 displays a scheme of the main decay paths after population of the $(^1L_a \pi\pi^*)$ and $(^1L_b \pi\pi^*)$ states at the FC region, together with the radiationless path mediated by $(^1L_b/gs)_{CI}$. The relative position of the latter CI with respect to the other singular points depicted in the figure supports the decrease of the measured fluorescence quantum yield with the increase of the excitation energy. After population of the bright $(^1L_a \pi\pi^*)$ state at the FC structure, the system has enough energy (~ 0.3 eV) to activate the nonradiative process via $(^1L_b/gs)_{CI}$, which is in agreement with relatively low fluorescence quantum yield observed by Glasser and Lami.¹⁴ If the indole molecule is irradiated with a light energy similar to the $(^1L_b \pi\pi^*)$ band origin, the $(^1L_b)_{min}$ region may be directly populated and the system will need an extra energy of 0.40 eV to access the $(^1L_b/gs)_{CI}$ funnel to the ground state. Therefore, the fluorescence from the $(^1L_b)_{min}$ will be the predominant process taking place. The nonradiative energy decay to the ground state will begin to compete with the emission of light as the irradiation energy approaches the $(^1L_a \pi\pi^*)$ vertical excitation energy, which is located 0.29 eV above the $(^1L_b/gs)_{CI}$ structure.

On the basis of our results, we associate therefore the nonradiative process observed at energies below 5.00 eV to be determined mainly by the above-described path from the $(^1L_b)_{min}$ minimum structure to the ground state via the $(^1L_b/gs)_{CI}$, which implies a threshold for the photophysical process around 0.5 eV lower with respect to the mechanism proposed by Sobolewski, Domcke, and co-workers^{23,26,27} and does not involve a dissociative $\pi\sigma^*$ state. It is worth mentioning however that the last energy decay route is still present to some extent, as it has been shown in some experiments in which the NH bond dissociation is observed,^{15,16} and it might also contribute at excitation energies lower than 5.00 eV via tunneling effects. In order to determine the percentage in which this dissociative deactivation path competes with the new mechanism obtained, and in particular, to ultimately understand the respective contribution of both energy decay paths to the photophysics of the system depending on the excited state populated and on the exceeding energy provided, a quantum dynamics study on bare indole would be necessary. Up to now only on-the-fly molecular

mechanic simulation based on time-dependent density functional theory (TDDFT) calculations starting on the S_2 state have been performed.⁴⁸ In such work, some trajectories computed using the PBE functional actually decay through the $\pi\sigma^*$ state, giving a low quantum yield for the dissociation along the NH coordinate. Nevertheless, according to the authors results and comments, the PBE functional is not able to provide the correct ordering of the $^1(L_a \pi\pi^*)$ and $^1(L_b \pi\pi^*)$ states and the use of the CAM-B3LYP functional, which on the contrary gives the correct ordering of the low-lying excited states, and provides a larger energy gap between the $\pi\pi^*$ and $\pi\sigma^*$ states, decreasing the probability of energy decay via the mechanism involving the NH bond stretch.

Triplet States. Along the evolution of the initially populated $^1(L_a \pi\pi^*)$ singlet state, and through the subsequent $^1(L_b \pi\pi^*)$ steepest descendent path leading to the $^1(L_b)_{\min}$ structure, no strongly efficient intersystem crossing regions have been identified. The same result has been obtained from the inspection of the two characterized secondary decay paths. In order to search for the phosphorescence structures related to the low-lying triplet excited states, $^3(L_a \pi\pi^*)$ and $^3(L_b \pi\pi^*)$, geometry optimizations of these states have been then carried out starting from the FC region. The minima obtained, namely $^3(L_a)_{\min}$ and $^3(L_b)_{\min}$, respectively, have an almost planar geometry (the corresponding structures can be found in Table S2, Supporting Information). The computed results, vertical excitation energies, band origins, and vertical emission energies are compiled in Table 2, together with the relevant experimental data. The $^3(L_a \pi\pi^*)$ state corresponds to the lowest triplet excited state at both the optimized $^3(L_a)_{\min}$ and $^3(L_b)_{\min}$ structures, and consequently the phosphorescence of the system is expected to take place from the $^3(L_a \pi\pi^*)$ state. In fact, a single phosphorescence band has been reported starting at 3.07 eV with a maximum at 2.87 eV,^{18,19} which is in agreement with previous²² and present theoretical findings for the $^3(L_a \pi\pi^*)$ state (see Table 2).

CONCLUSIONS

On the basis of the exploration of the PEH on the low-lying singlet and triplet excited states by means of accurate state-of-the-art computational strategies, a global and comprehensive description of the photophysics of bare indole under UV irradiation is provided. Two types of processes have been identified and characterized in the molecule after excitation to the low-lying singlet valence excited states $^1(L_a \pi\pi^*)$ and $^1(L_b \pi\pi^*)$: radiative and nonradiative energy decays. After UV absorption most of the molecules reach the bright $^1(L_a \pi\pi^*)$ excited state, which would evolve through a steepest descendent path toward a CI implying the $^1(L_b \pi\pi^*)$ state, mediating in this manner the energy transfer to the latter state. The relaxed structure $^1(L_b)_{\min}$ on the S_1 hypersurface might be accessed subsequently and from this region the fluorescence emission takes place. All the secondary energy decay paths studied also lead to the emissive state, $^1(L_b)_{\min}$. A radiationless decay specific to the $^1(L_b \pi\pi^*)$ state has been predicted involving a nonplanar ethene-like CI between $^1(L_b \pi\pi^*)$ and the ground state. According to the obtained pathways profiles, the process is inhibited at the energy around the $^1(L_b \pi\pi^*)$ band origin and becomes competitive with the fluorescence emission at excitation energies close the $^1(L_a \pi\pi^*)$ band maximum. The current findings explain for the first time the low fluorescence quantum yield measured

experimentally after excitation to energy lower than 5.00 eV because of the presence of a not previously reported nonradiative decay mechanism specific for the $^1(L_b \pi\pi^*)$ state. No efficient intersystem crossings have been found along the decay paths analyzed, while the phosphorescence properties are in agreement with previous theoretical and experimental studies.

Two additional comments are appropriate at this final stage. First, as many other organic molecules, the nonradiative decay to the ground state of excited indole is mediated by the presence of an ethene-like CI. For instance, the photostability of the natural purines and pyrimidines nucleobases has been proven to be determined by the presence of easily accessible ethene-like CI, which provides the molecules with nonradiative paths to release the absorbed energy in an ultrafast manner.^{31,32,49–54} The present study on the photophysics of bare indole further highlights the importance of such type of degenerate regions, which can be considered nowadays as a well-established concept in the field of efficient nonradiative decay funnels of organic systems. Second, since the topology of indole and the canonical nucleobase 9H-adenine are similar, that is, they are both built as six- and five-fused rings and the molecules have the same energetic ordering of the lowest $^1(L_b \pi\pi^*)$ and $^1(L_a \pi\pi^*)$ states at the FC region, the photophysics of 9H-adenine^{49–53} might be compared with the results obtained here. The presence of a main nonradiative decay path toward the ground state, after excitation to the bright state, associated with an ethene-like CI appears as a common feature between the two systems, although involving the $^1(L_b \pi\pi^*)$ and $^1(L_a \pi\pi^*)$ states in indole and 9H-adenine, respectively. This can be ascribed to the different nature of the double bond implied in the CI of indole and 9H-adenine ($C_6=C_7$ versus $C_2=N_3$) and can consequently be related to the intrinsic nature of the two molecules.

ASSOCIATED CONTENT

S Supporting Information. The natural orbitals of the active space employed in the CASSCF/CASPT2 calculations are depicted in Figure S1. Table S1 compiles the CASPT2 E_{VA} of indole at the FC geometry computed with two different basis sets including or not Rydberg functions. The CASSCF MEPs related to Figures 3–6, and corresponding to the $^1(L_a \pi\pi^*)$ MEP from the FC structure, $^1(L_b \pi\pi^*)$ MEPs from $(^1L_a/^1L_b)_{CI\alpha}$ and the FC geometry, $^1(L_a \pi\pi^*)$ MEP from $(^1L_a/^1L_b)_{CI\beta}$, and $^1(L_b \pi\pi^*)$ MEPs from $(^1L_a/^1L_b)_{CI\beta}$ and $(^1L_b/gs)_{CI}$ are displayed in Figures S2–7, respectively. The x, y, z coordinates for the structures of the present study can be found in Table S2. This material is available free of charge via the Internet at <http://pubs.acs.org>.

AUTHOR INFORMATION

Corresponding Author

*E-mail: Angelo.Giussani@uv.es; Daniel.Roca@kvac.uu.se.

ACKNOWLEDGMENT

Research supported by the project CTQ2010-14892 of the Spanish MEC/FEDER and the Swedish Research Council. A.G. gratefully acknowledges Ph.D. fellowship “V segles” from the Universitat de València. A.G. also thanks Prof. N. Ferré for useful discussion. D.R.-S. thanks European Research Council under the European Community’s Seventh Framework Programme (FP7/2007-2013)/ERC grant agreement 255363.

REFERENCES

- (1) Wetlaufer, D. B. Ultraviolet spectra of proteins and amino acids. In *Advances in Protein Chemistry*; Anfinsen, C. B., Anson, M. L., Bailey, K., Edsall, J. T., Eds.; Academic Press: New York, 1962; Vol. 17.
- (2) Demchenko, A. P. *Ultraviolet Spectroscopy of Proteins*; Springer-Verlag: Berlin, 1986.
- (3) Creed, D. *Photochem. Photobiol.* **1984**, *39*, 537.
- (4) Hollas, J. M. *Spectrochim. Acta* **1963**, *19*, 753.
- (5) Lami, H. *J. Chem. Phys.* **1977**, *67*, 3274.
- (6) Lami, H. *Chem. Phys. Lett.* **1977**, *48*, 447.
- (7) Ilich, P. *Can. J. Spectrosc.* **1987**, *67*, 3274.
- (8) Bartis, T. L. O.; Grace, L. I.; Dunn, T. M.; Lubman, D. M. *J. Phys. Chem.* **1993**, *97*, 5820.
- (9) Strickland, E. H.; Horwitz, J.; Billups, C. *Biochemistry* **1970**, *25*, 4914.
- (10) Fender, B. J.; Sammeth, D. M.; Callis, P. R. *Chem. Phys. Lett.* **1995**, *239*, 31.
- (11) Lami, H.; Glasser, N. *J. Chem. Phys.* **1986**, *84*, 597.
- (12) Chang, C. T.; Wu, C. Y.; Muirhead, A. R.; Lombardi, J. R. *Photochem. Photobiol.* **1974**, *19*, 347.
- (13) Caminati, W.; Di Bernardo, S. *J. Mol. Struct.* **1990**, *240*, 253.
- (14) Glasser, N.; Lami, H. *J. Chem. Phys.* **1981**, *74*, 6526.
- (15) Lin, M. F.; Tseng, C. M.; Lee, Y. T.; Ni, C. K. *J. Chem. Phys.* **2005**, *123*, 124303.
- (16) Nix, M. G. D.; Devine, A. L.; Cronin, B.; Ashfold, M. N. R. *Phys. Chem. Chem. Phys.* **2006**, *8*, 2610.
- (17) Borisevich, N. A.; Sukhodola, A. A.; Tolstorozhev, G. B. *Chem. Phys.* **2008**, *354*, 44.
- (18) Zuckich, J.; von Schütz, J. U.; Maki, A. M. *J. Am. Chem. Soc.* **1974**, *96*, 710.
- (19) Wilkinson, F.; Garmer, A. *J. Chem. Soc., Faraday Trans. 2* **1977**, *73*, 222.
- (20) Brand, C.; Küpper, J.; Pratt, D. W.; Meerts, W. L.; Krügler, D.; Tatchen, J.; Schmitt, M. *Phys. Chem. Chem. Phys.* **2010**, *12*, 4968.
- (21) Küpper, J.; Pratt, D. W.; Meerts, W. L.; Brand, C.; Tatchen, J.; Schmitt, M. *Phys. Chem. Chem. Phys.* **2010**, *12*, 4980.
- (22) Serrano-Andrés, L.; Roos, B. O. *J. Am. Chem. Soc.* **1996**, *118*, 185.
- (23) Sobolewski, A. L.; Domcke, W. *Chem. Phys. Lett.* **1999**, *315*, 293.
- (24) Borin, A. C.; Serrano-Andrés, L. *Chem. Phys.* **2000**, *262*, 253.
- (25) Serrano-Andrés, L.; Borin, A. C. *Chem. Phys.* **2000**, *262*, 267.
- (26) Sobolewski, A. L.; Domcke, W.; Dedonder-Lardeux, C.; Jouvet, C. *Phys. Chem. Chem. Phys.* **2002**, *4*, 1093.
- (27) Sobolewski, A. L.; Domcke, W. *J. Phys. Chem. A* **2007**, *111*, 11725.
- (28) Platt, J. R. *J. Chem. Phys.* **1949**, *17*, 489.
- (29) De Vico, L.; Olivucci, M.; Lindh, R. *J. Chem. Theory Comput.* **2005**, *1*, 1029.
- (30) Anglada, J. M.; Bofill, J. M. *J. Comput. Chem.* **1997**, *18*, 992.
- (31) Merchán, M.; González-Luque, R.; Climent, T.; Serrano-Andrés, L.; Rodríguez, E.; Reguero, M.; Peláez, D. *J. Phys. Chem. B* **2006**, *110*, 26471.
- (32) Serrano-Andrés, L.; Merchán, M.; Borin, A. C. *J. Am. Chem. Soc.* **2008**, *130*, 2473.
- (33) Andersson, K.; Malmqvist, P.-Å.; Roos, B. O. *J. Chem. Phys.* **1992**, *96*, 1218.
- (34) Roos, B. O.; Andersson, K.; Fülcher, M. P.; Malmqvist, P.-Å.; Serrano-Andrés, L.; Pierloot, K.; Merchán, M. *Adv. Chem. Phys.* **1996**, *93*, 219.
- (35) Serrano-Andrés, L.; Merchán, M. Spectroscopy: Applications. In *Encyclopedia of Computational Chemistry*; Schleyer, P. v. R., Schreiner, P. R., Schaefer, H. F., III, Jorgensen, W. L., Thiel, W., Glen, R. C., Eds.; Wiley, Chichester, 2004; pp 1–51.
- (36) Roca-Sanjuán, D.; Aquilante, F.; Lindh, R. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2011**; DOI: 10.1002/wcms.97.
- (37) Aquilante, F.; De Vico, L.; Ferré, N.; Ghigo, G.; Malmqvist, P.-Å.; Pedersen, T.; Pitonak, M.; Reiher, M.; Roos, B. O.; Serrano-Andrés, L.; Urban, M.; Varyazov, V.; Lindh, R. *J. Comput. Chem.* **2010**, *31*, 224.
- (38) Merchán, M.; Serrano-Andrés, L. Ab Initio Method for Excited States. In *Computational Photochemistry*, 1st ed.; Olivucci, M., Ed.; Elsevier: Amsterdam, The Netherlands, 2005; Vol. 16, pp 35–91.
- (39) Serrano-Andrés, L.; Merchán, M. Photostability and Photo-reactivity in Biomolecules: Quantum Chemistry of Nucleic Acid Base Monomers and Dimers. In *Radiation Induced Molecular Phenomena in Nucleic Acids*; Leszczynski, J., Shukla, M., Eds.; Springer: Heidelberg, The Netherlands, 2008; pp 435–472.
- (40) Garavelli, M. *Theor. Chem. Acc.* **2006**, *116*, 87.
- (41) Garavelli, M.; Bernardi, F.; Cembran, A. Ab Initio Method for Excited States. In *Computational Photochemistry*, 1rd ed.; Olivucci, M., Ed.; Elsevier: Amsterdam, The Netherlands, 2005; Vol. 16, pp 191–223.
- (42) Takigawa, T.; Ashida, T.; Sasado, Y.; Kakuda, M. *Bull. Chem. Soc. Jpn.* **1966**, *39*, 2369.
- (43) Michl, J. In *Computational Photochemistry*, 1rd ed.; Olivucci, M., Ed.; Elsevier: Amsterdam, The Netherlands, 2005; Vol. 16, pp ix–xi.
- (44) Song, P. S.; Kurtin, W. E. *J. Am. Chem. Soc.* **1969**, *91*, 4892.
- (45) Eftink, M. R. Fluorescence techniques for studying protein structure. In *Methods of Biochemical Analysis*; Suelter, C. H., Ed.; John Wiley & Sons: New York, 1991; Vol. 35.
- (46) Garavelli, M.; Celani, P.; Yamamoto, N.; Bernardi, F.; Robb, M. A.; Olivucci, M. *J. Am. Chem. Soc.* **1996**, *118*, 11656.
- (47) Garavelli, M.; Bernardi, F.; Cembran, A.; Castaño, O.; Frutos, L. M.; Merchán, M.; Olivucci, M. *J. Am. Chem. Soc.* **2002**, *124*, 13770.
- (48) Wohlgemuth, M.; Bona i-Koutecký, V.; Mitrić, R. *J. Chem. Phys.* **2011**, *135*, 054105.
- (49) Perun, S.; Sobolewski, A. L.; Domcke, W. *J. Am. Chem. Soc.* **2004**, *127*, 6257.
- (50) Blancafort, L. *J. Am. Chem. Soc.* **2006**, *128*, 210.
- (51) Serrano-Andrés, L.; Merchán, M.; Borin, A. C. *Chem.—Eur. J.* **2006**, *12*, 6559.
- (52) Serrano-Andrés, L.; Merchán, M.; Borin, A. C. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 8691.
- (53) Conti, I.; Garavelli, M.; Orlandi, G. *J. Am. Chem. Soc.* **2009**, *131*, 16108.
- (54) Gobbo, J. P.; Borin, A. C.; Serrano-Andrés, L. *J. Phys. Chem. B* **2011**, *115*, 6243.

Direct Calculations of Mid- and Near-IR Absorption and Circular Dichroism Spectra of Chiral Molecules Using QM/MM Molecular Dynamics Simulation Method

Jun-Ho Choi[†] and Minhaeng Cho^{*,†,‡}

[†]Department of Chemistry, Korea University, Seoul 136-701, Korea

[‡]Multidimensional Spectroscopy Laboratory, Korea Basic Science Institute, Seoul 1360-713, Korea

ABSTRACT: The infrared (IR) and vibrational circular dichroism (VCD) spectra of (1*S*)-(–)- β -pinene in the mid- and near-IR frequency regions are numerically simulated by using a time-correlation function theory and mixed quantum/classical simulation method. Anharmonic vibrational dynamics and fluctuating atomic partial charges of the chiral pinene were obtained by carrying out quantum mechanical/molecular mechanical (QM/MM) molecular dynamics (MD) simulations. Thus obtained time-correlation functions of electric and magnetic dipole moments are used to calculate the IR absorption and VCD spectra, and they are directly compared with experimental results. Not only the fundamental transition bands but also first overtone and combination bands in the near-IR frequency region are successfully simulated. It is shown that the polarizable nature of the solute is particularly important in quantitatively reproducing the near-IR spectra, whereas such polarization effects on dipole and rotational strengths of lower-frequency and large-amplitude vibrations are less critical. We anticipate that the present QM/MM MD method in combination with mixed quantum/classical time-correlation function theory to calculate both mid- and near-IR absorption and VCD spectra will be of critical use in interpreting vibrational optical activity properties of even conformationally flexible chiral molecules, such as proteins.

1. INTRODUCTION

Near-infrared (NIR) spectroscopy uses near-infrared region of the electromagnetic field in the wavelength range from 800 nm to 2.5 μm .¹ Within this wavelength range, molecular overtone and combination vibrations can be resonantly excited.^{2,3} Such transitions are, however, quantum mechanically forbidden within the harmonic approximations to molecular vibrations so that their oscillator strengths are often 1 or 2 orders of magnitude smaller than those of fundamental transitions, which makes it difficult to use as a sensitive technique. However, a notable advantage of NIR spectroscopy is its ability to penetrate farther into a sample than mid-infrared or visible radiation. Thereby, it has been used in the analyses of blood hemoglobin levels, pharmaceuticals, combustion products, and even atmospheres of cool stars.^{4–7}

Typically, the molecular overtone and combination absorption bands in such NIR regions are quite broad and complicated with multiple peaks. Therefore, assigning specific spectral features to particular quantum transitions and chemical components has been extremely challenging.^{8,9} To enhance the frequency resolution of the NIR transition bands and to extract detailed information on absolute configuration of chiral molecule, the NIR vibrational circular dichroism (NIR-VCD) spectroscopy has been used.^{10–22} In 1976, Keiderling and Stephens for the first time measured the NIR-VCD spectrum of the C–H stretching overtone and C–H stretching/C–H bending combination bands of camphor dissolved in CCl_4 solution.¹⁶ Recently, Nafie and Abbate groups have reported the NIR-VCD spectra of chiral organic molecules including terpenes.^{11,19,21–23} The C–H stretching first and second overtones, C=O stretching second overtone, and various combination modes of C–H stretching and bending, C–C stretching, C=C stretching, and O–H

stretching and bending have been observed in the NIR absorption and VCD spectra of which frequency range is from 4000 to 10 000 cm^{-1} .¹¹

In parallel with these experimental efforts, a number of novel attempts to numerically simulate the NIR absorption and VCD spectra by using ab initio quantum chemistry calculation methods have been reported recently.^{20,24–29} For example, Abbate and co-workers used the local mode approximation,^{8,9,30–32} which was initially developed by Child and Henry, for C–H stretching vibrations of camphor and camphorquinone and presented the simulated NIR-VCD spectra of C–H stretching overtone ($\Delta\nu = 3$) transition at $\sim 8500 \text{ cm}^{-1}$.³³ The C–H stretch local mode was treated as a Morse oscillator³⁴ so that the matrix elements of coordinate and momentum operators could be calculated by using analytical expressions for the vibrational wave functions of the Morse oscillator. The coordinate-dependent atomic polar and axial tensors that are needed to determine dipole and rotational strengths of vibrational excitations were calculated by using a density function theory method, i.e., B3LYP/6-31G**. Their calculation method for simulating C–H stretching NIR spectra was shown to be quite successful.³³ However, such approaches may not be easily extended to simulate the NIR spectra for a variety of combination bands other than C–H stretching overtones as well as those of conformationally flexible molecular systems, since the complicated electric and mechanical anharmonicity effects on transition frequencies and electric and magnetic dipole moments are increasingly difficult to take into consideration by using such quantum chemistry calculation method for an isolated chiral molecule.^{10,35}

Received: September 20, 2011

Published: November 10, 2011

In addition, the solvent–solute interaction-induced polarization and line-broadening effects on the NIR absorption and VCD spectra could not be taken into account by considering isolated molecules.

In order to overcome these limitations of *ab initio* calculation approaches, molecular dynamics (MD) simulation method combined with an electronic structure calculation method has recently been used to simulate IR and VCD spectra of chiral molecules in solutions.^{36–46} For instance, Yang and Cho numerically simulated the IR absorption spectrum of *N*-methylacetamide (NMA) in water, directly calculating the electric dipole autocorrelation function of the NMA that is treated with a semiempirical quantum mechanical (QM) method and analyzing the QM/MM MD trajectories.⁴¹ Later, a Hartree–Fock (HF)-QM/MM MD simulation method was used to simulate both IR absorption and VCD spectra of blocked alanine (Ac-Ala-NHMe) in water by carrying out Fourier transforms of the correlation functions of electric and magnetic dipole moments.⁴⁰ Such a QM/MM MD simulation method combined with time-correlation function theory for IR absorption and VCD spectroscopy was shown to be quite useful even for molecular systems with shallow potential energy surfaces with thermally accessible multiple conformations.^{40,47,48} The fundamental transition bands thus obtained were in fair agreement with experimental results.⁴⁹ Here, it should be emphasized that the present approach with QM/MM MD simulation method combined with time-correlation function theory has an advantage in taking into account fluctuating potential anharmonicity and its effects on mode couplings, vibrational frequencies, and forbidden transition moments, since the on-the-fly QM calculations of the solute correctly include all the possible anharmonic effects that cannot be correctly taken into account with fixed charge MM simulations. Thus, it is not necessary to use a theoretical model based on local mode basis nor to carry out approximate finite-difference calculations of a variety of mechanical and electric anharmonic properties. Other notable examples demonstrating the use of QM/MM MD simulation methods in the molecular spectroscopy are needed to be mentioned here. Ringer and MacKerell calculated the vibrational frequency shifts of the CN stretching mode of acetonitrile in two different solvent environments.⁵⁰ Corcelli and co-workers considered C–D stretching mode of Ala-d₁ in an aqueous solution and successfully showed that the multiple absorption peaks found in the 2000–2300 cm⁻¹ region in fact result from Fermi resonances with fundamental C_α–D stretching vibration.³⁸

In this paper, we shall present QM/MM MD simulation results of the (1*S*)-(–)-β-pinene in CCl₄ solution to simulate both mid- and near-IR absorption and VCD spectra. This is the first attempt to directly calculate such NIR spectra by using time-correlation function approaches combined with a mixed QM/MM method. The simulated NIR absorption and VCD spectra exhibiting both C–H stretching first overtone and various C–H stretching/C–H bending combination bands of (1*S*)-(–)-β-pinene are directly compared with experimental results reported by Guo et al. in ref 11.

2. COMPUTATIONAL METHOD

The molecular structure of (1*S*)-(–)-β-pinene is shown in Figure 1. This pinene was treated quantum mechanically with HF/6-31G method in the present QM/MM MD simulation studies, where 186 CCl₄ solvent molecules were treated classical

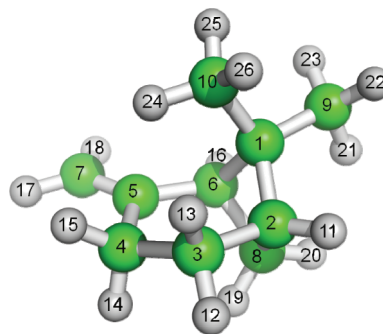


Figure 1. Molecular structure of (1*S*)-(–)-β-pinene. The atom types are also shown in this figure (see Table 1 for average Mulliken partial charges and standard deviations).

mechanically with CHARMM⁵¹ force fields (version c33b2). To carry out the QM/MM MD simulation, we used the CHARMM program interfaced with GAMESS-UK⁵² electronic structure package (version 7.0). A single solute molecule was placed at the center of a spherical solvent cluster with radius of 25 Å. A restraining harmonic potential was used to prevent evaporation of solvent molecules at 273 K. The solute–solvent cluster system was first minimized with the steepest descent and adopted basis Newton–Raphson methods for 100 steps each, and then the system was heated to 273 K and equilibrated for 25 ps. Simulation time step used was 0.5 fs, and totally 1 ns production run was performed. Cartesian coordinates, velocities, and atomic partial charges of the solute were saved for every 1.0 fs for subsequent calculations of electric and magnetic dipole moment trajectories. The spectral range of the simulated vibrational spectra is determined by the Nyquist critical frequency $f_c = 1/(2\Delta_t)$, where Δ_t corresponds to the data saving time step of 1 fs.

Now, using the QM/MM MD trajectory with saved atomic Mulliken partial charges of the solute, we could obtain the fluctuating electric and magnetic dipole moments, denoted as $\boldsymbol{\mu}(t)$ and $\mathbf{M}(t)$, respectively, where they are defined as

$$\begin{aligned}\boldsymbol{\mu}(t) &= \sum_i q_i(t)\mathbf{r}_i(t) \\ \mathbf{M}(t) &= \frac{1}{2c} \sum_i q_i(t)\mathbf{r}_i(t) \times \mathbf{v}_i(t)\end{aligned}\quad (1)$$

Here, $q_i(t)$, $\mathbf{r}_i(t)$, and $\mathbf{v}_i(t)$ represent the partial charge,⁵³ position, and velocity of the *i*th solute atom. In this work, we used the atomic Mulliken partial charges instead of Löwdin charges, because the numerically simulated VCD spectrum of the C–H stretching vibrations with the former charges is found to be in better agreement with the experimentally measured spectrum. The atomic partial charges explicitly depend on time due to solute–solvent interaction-induced charge response and conformational fluctuation of the solute molecule. Thereby, such polarizable nature of the solute molecule is naturally taken into account by the present QM/MM MD simulation method. Using thus obtained electric and magnetic dipole moment trajectories, one can readily calculate the autocorrelation function of the electric dipole moment as well as the cross-correlation function between the electric and magnetic dipole moments. Then, the IR absorption ($A(\omega)$) and VCD ($\Delta A(\omega)$) spectra can be directly calculated by carrying out the following Fourier transformations of the corresponding time-correlation

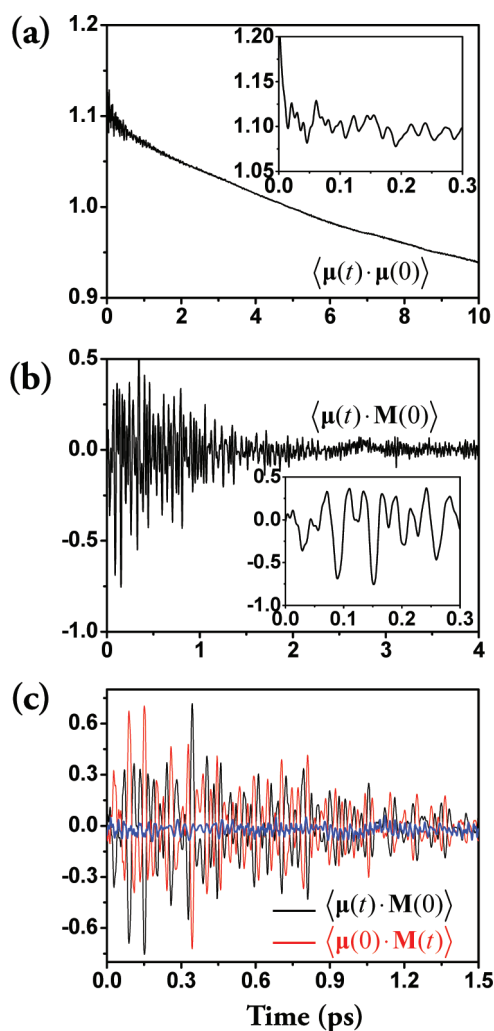


Figure 2. Time-correlation functions $\langle \mu(t) \cdot \mu(0) \rangle$ and $\langle \mu(t) \cdot \mathbf{M}(0) \rangle$ are shown in (a) and (b), respectively. The dimensions of $\langle \mu(t) \cdot \mu(0) \rangle$ and $\langle \mu(t) \cdot \mathbf{M}(0) \rangle$ are in D^2 and $10^{-4} D\mu_B$, where D and μ_B are debye and bohr magneton. The first 300 fs portions of $\langle \mu(t) \cdot \mu(0) \rangle$ and $\langle \mu(t) \cdot \mathbf{M}(0) \rangle$ are shown in the inset of (a) and (b). The two correlation functions, $\langle \mu(t) \cdot \mathbf{M}(0) \rangle$ (black line) and $\langle \mu(0) \cdot \mathbf{M}(t) \rangle$ (red line), which are to be mirror images in an ideal case, are plotted in (c). The sum of the two is the blue line in (c).

functions, i.e.,^{40,54}

$$A(\omega) \propto \int_{-\infty}^{\infty} dt e^{i\omega t} \langle \mu(t) \cdot \mu(0) \rangle \quad (2)$$

$$\Delta A(\omega) \equiv A_L(\omega) - A_R(\omega) \propto \text{Im} \int_{-\infty}^{\infty} dt e^{i\omega t} \langle \mu(t) \cdot \mathbf{M}(0) \rangle \quad (3)$$

where $A_L(\omega)$ and $A_R(\omega)$ are the absorption spectrum measured by using a left and right circularly polarized radiation, respectively. A detailed theoretical description on the time-correlation function formalism of the VCD spectrum can be found in ref 40.

3. SIMULATED IR ABSORPTION AND VCD SPECTRA

3.1. Time-Correlation Functions. The two time-correlation functions, $\langle \mu(t) \cdot \mu(0) \rangle$ and $\langle \mu(t) \cdot \mathbf{M}(0) \rangle$, calculated with eq 1 are shown in Figure 2. It turns out that, even though the overall

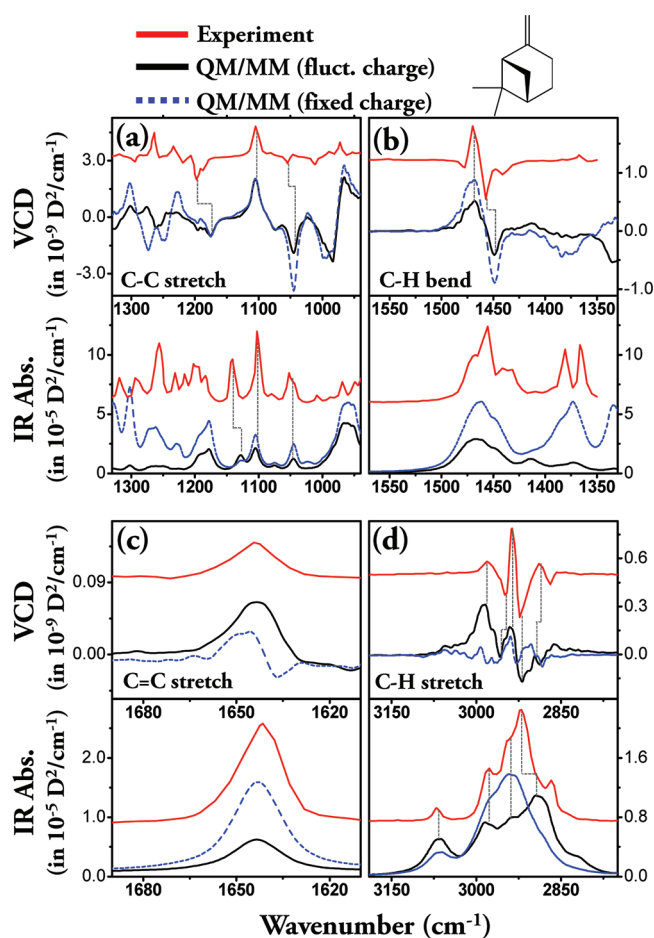


Figure 3. Numerically simulated IR and VCD spectra of (1S)-(-)- β -pinene. Those of C–C stretch (a), C–H bend (b), C=C stretch (c), and C–H stretch (d) modes are shown in this figure. The red lines represent experimentally measured IR absorption and VCD spectra reported in ref 11. The black lines represent the simulated spectra obtained by using the QM/MM MD simulation trajectory and eq 1 so that the fluctuating charges were properly taken into account. The blue lines however correspond to the simulated spectra with fixed atomic partial charges (see Table 1). Here, the frequency correction factors for the C–C stretch, C–H bend, C=C stretch, and C–H stretch modes are 0.8701, 0.8739, 0.8779, and 0.8953, respectively. Also the corresponding lifetimes are assumed to be 1 ps and 800, 600, and 500 fs. Note that the IR absorption and VCD intensities are given in 10^{-5} and 10^{-9} D^2/cm^{-1} , respectively.

absorption and VCD line shapes are clearly discernible, the direct Fourier transforms of these correlation functions appear to be quite noisy due to the limited length of QM/MM MD trajectory. Furthermore, the intrinsic lifetime-broadening contribution to the simulated spectra could not be taken into account, since the QM/MM MD simulations were performed at a thermal equilibrium condition. Therefore, we approximately describe the lifetime effect by multiplying an exponential function of $\exp(-t/2T_1)$ to the time-correlation function before performing Fourier transformations in eqs 2 and 3, where T_1 represents the excited-state lifetime of a given mode and is treated as an adjustable parameter. In Figure 2a, the original $\mu-\mu$ correlation function without lifetime-broadening factor is plotted. In addition to a slowly decaying component reflecting rotational relaxation of the pinene, there are highly oscillating features, as

can be seen in the first 300 fs period of the correlation function (see the inset of Figure 2a). These oscillating components contain information about mid- and near-IR-active mode frequencies. In Figure 2b, the μ - M cross-correlation function is plotted, and its short time (up to 300 fs) region is shown in the inset. Again, the seemingly irregular oscillating pattern in the cross-correlation function is essentially produced by a superposition of various cross-correlation functions associated with intramolecular vibrational dynamics dictated by intrinsically multidimensional anharmonic potential energy surface. Before the Fourier-transformed IR and VCD spectra are presented and discussed, an important convergence issue should be briefly discussed. If the trajectory is sufficiently long and the ensemble-averaged μ - M cross-correlation is statistically acceptable, it is expected that $\langle \mu(t) \cdot M(0) \rangle = -\langle \mu(0) \cdot M(t) \rangle$.^{40,54} In Figure 2c, these two cross-correlation functions are plotted for the sake of comparison. The fact that the sum (blue line) of the two shown in the same figure is very small indicates that the MD trajectory is sufficiently long, and the ensemble average required to calculate the correlation function is reasonably good.

3.2. Fourier-Transformed IR Absorption and VCD Spectra.

Using the μ - μ and μ - M correlation functions, one can readily calculate their Fourier-transforms that correspond to IR absorption and VCD spectra, after treating the lifetime broadening in an ad hoc manner (see the caption of Figure 3 for mode-dependent lifetimes used here). In Figure 3, the fundamental transition bands in the frequency range from 800 to 3300 cm^{-1} are shown (see black lines), where the red lines are the experimentally measured spectra that can be found in ref 11. The IR peak intensities of fundamental transition modes vary from 1×10^{-5} to $4 \times 10^{-5} \text{ D}^2/\text{cm}^{-1}$, whereas the corresponding VCD intensities are in the range from -2×10^{-9} to $2 \times 10^{-9} \text{ D}^2/\text{cm}^{-1}$. Therefore, the ratio $\Delta A/A$ is estimated to be about 10^{-4} , which is fully consistent with the experimental results. The delocalized C-C stretching (800–1350 cm^{-1}), C-H bending (1350–1550 cm^{-1}), C=C stretching (1600–1700 cm^{-1}), and C-H stretching (2850–3150 cm^{-1}) bands are separately plotted in Figure 3a–d, respectively. Usually, a single scaling factor was used to correct all the ab initio calculated harmonic normal-mode frequencies of an isolated molecule—note that, for HF/6-31G method, the corresponding scaling factor used quite often is 0.8953.⁵⁵ However, for the sake of comparisons with experimental results, we use slightly different scaling factors for the four different groups of normal modes (see the caption of Figure 3). In the cases of the C-C stretching bands, the peak positions in the simulated IR absorption and VCD spectra cannot be directly matched to those in the experimental spectra. Nevertheless, a few strong features, such as the intense positive VCD peak at $\sim 1100 \text{ cm}^{-1}$ and the negative VCD peak at 1050 cm^{-1} , are well reproduced by the QM/MM MD simulation.

The agreement between theory and experiment is excellent for the C-H bending IR and VCD bands appearing in the frequency range from 1425 to 1500 cm^{-1} (Figure 3b). The negative–positive (from low- to high-frequency region) couplet feature in the corresponding VCD spectrum is clearly visible. The IR absorption and VCD spectra of C=C stretching vibration are also in quantitative agreement with experimental results (see Figure 3c). However, the agreement between the simulated C-H stretching IR and VCD spectra and the experimental results is less quantitative in comparison to those of C-H bending and C=C stretching bands. Nonetheless, the overall VCD spectral pattern, i.e., “W”-form line shape, is nicely reproduced by the

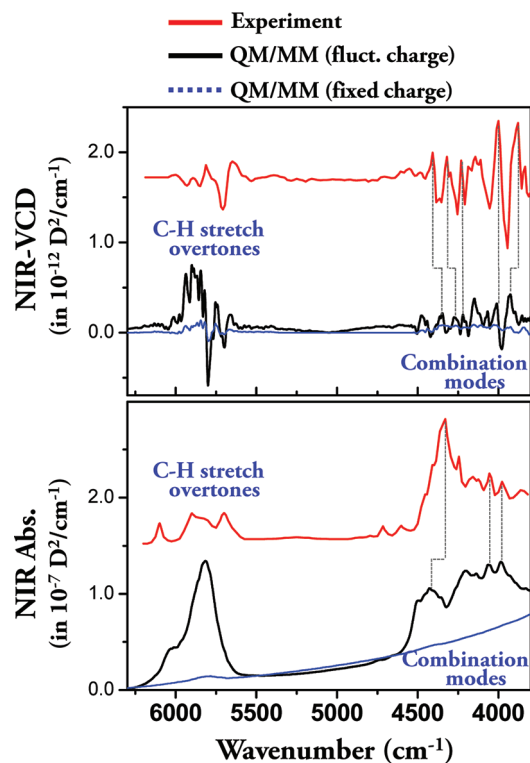


Figure 4. Numerically simulated near-IR absorption and VCD spectra. The red lines represent experimentally measured IR absorption and VCD spectra reported in ref 11. The black lines are the simulated spectra obtained by using the QM/MM MD trajectory and eq 1. The blue lines are the simulated spectra with fixed atomic partial charges. The frequency correction factor for these C-H overtone and combination modes is 0.8953, and the lifetimes are assumed to be 300 fs. Note that the NIR absorption and NIR-VCD intensities are given in 10^{-7} and $10^{-12} \text{ D}^2/\text{cm}^{-1}$, respectively.

present simulation method. In summary, we find that: (1) the numerically simulated IR and VCD spectra of highly delocalized low-frequency C-C backbone stretching vibrations are still difficult to be accurately reproduced by the present QM/MM MD simulation method in combination with classical time-correlation function approach, (2) the line shapes of C-H bend and C=C stretch IR and VCD spectra are successfully simulated, and (3) the W-shape spectral feature of the C-H stretch VCD spectrum is in fair agreement with experimental result.

3.3. Near-IR Absorption and VCD Spectra. An important advantage of the present time-correlation function approach is that the direct Fourier transform of the corresponding correlation function provides both mid- and near-IR spectra simultaneously, as long as the time step for numerically calculated correlation function is sufficiently small. The NIR region of the vibrational spectrum contains information on resonant frequencies of a variety of overtone and combination modes. In the present case of the pinene, there are two NIR bands in the frequency range from 3800 to 6100 cm^{-1} , as can be seen in Figure 4. Again, the experimentally measured IR and VCD spectra (red lines) are given in the figure for the sake of direct comparisons. Here, the NIR absorption peak intensities of both combination and the overtone transition modes are in the range from 0.4×10^{-9} to $1.8 \times 10^{-9} \text{ D}^2/\text{cm}^{-1}$, whereas the corresponding NIR-VCD peak intensities vary from -2×10^{-12} to

$2 \times 10^{-12} \text{ D}^2/\text{cm}^{-1}$. Therefore, the ratio $\Delta A/A$ in the NIR frequency region is about 10^{-4} or 10^{-5} , which is again consistent with the experimental results. Furthermore, the relative IR or VCD intensities of fundamental transitions modes are found to be 2–3 orders of magnitude larger than those of combination or overtone transition modes (compare the y -axis scales of Figure 4 with those of Figure 3).

In the NIR absorption spectrum, the overtone transition band of C–H stretching vibrations appears in a high-frequency region ($5500\text{--}6100 \text{ cm}^{-1}$). Despite that the experimental spectrum exhibits three peaks, our spectrum appears to be broad and featureless except for a shoulder band in the high-frequency side. Now, the lower frequency band ($3800\text{--}4700 \text{ cm}^{-1}$) originates from transitions of various combination modes. In ref 11 it was shown that the combination band consists of three sub-bands, i.e., combination of C–H and C–C stretch ($3800\text{--}4200 \text{ cm}^{-1}$), combination of C–H stretch and C–H bend ($4200\text{--}4500 \text{ cm}^{-1}$), and combination of C–H (in C=CH group) and C=C stretch ($4500\text{--}4700 \text{ cm}^{-1}$). Despite that there exist certain discrepancies in the spectral intensity distribution between the theoretical and experimental spectra, the present time-correlation function approach utilizing HF/6-31G/MM MD trajectory correctly provides information on the peak positions of combination and the first C–H stretching overtone bands.

We next consider the numerically simulated C–H overtone and combination VCD spectra shown in the upper panel of Figure 4. Although the positive–negative alternating features in the combination band ($3800\text{--}4700 \text{ cm}^{-1}$) are fairly well reproduced by the simulated spectrum, the overall intensity distribution and precise peak positions are not. Furthermore, the peak pattern of the C–H overtone band found in the experimentally measured VCD spectrum is not quantitatively reproduced by the simulated spectrum. Note that the VCD spectral pattern results from complicated superposition of a number of positive and negative features. Consequently, any inaccuracy in predicting frequencies of anharmonic vibrational modes would lead to significantly distorted spectrum deviating from experiment. However, even though typical IR absorption and VCD intensities of the combination and overtone bands are 1–2 orders of magnitude smaller than the fundamental transition bands, such weak features are clearly visible in the simulated NIR absorption and VCD spectra obtained from the calculated $\mu\text{--}\mu$ and $\mu\text{--}\mathbf{M}$ time-correlation functions. This strongly indicates that the molecular dynamics of the pinene on the QM anharmonic potential energy surface is clearly better than that on the classical mechanical potential energy surface constructed by MM force fields.

3.4. Fluctuating Charge Effects on NIR Spectra. As can be seen in eq 1, the fluctuating electric and magnetic dipoles are determined by time-dependent atomic partial charges, coordinates and velocities. In the case of the MM MD simulation, the atomic partial charges are often assumed to be constants so that the polarizable nature of the solute molecule is completely ignored. However, in the present QM/MM MD simulation, fluctuating atomic partial charges due to solute–solvent interactions were constantly updated and saved for subsequent calculations of electric and magnetic dipole moments. Now, to investigate the precise effect of the solute polarization on vibrational spectra, we deliberately consider the average atomic partial charges, instead of time-dependent ones, when the electric and magnetic dipole moments are calculated, i.e., $\mu(t) \cong \sum_i \langle q_i \rangle \mathbf{r}_i(t)$

Table 1. Average Atomic Partial Charges and Their Standard Deviations in Parentheses of (1S)-(–)- β -Pinene Obtained From QM/MM MD Simulation Trajectory^a

atom type	atomic partial charge (in unit e)
C1	–0.017 (0.014)
C2	–0.210 (0.010)
C3	–0.281 (0.011)
C4	–0.373 (0.017)
C5	0.165 (0.019)
C6	–0.295 (0.016)
C7	–0.424 (0.016)
C8	–0.272 (0.011)
C9	–0.435 (0.013)
C10	–0.446 (0.016)
H11 (-C2)	0.160 (0.008)
H12 (-C3)	0.161 (0.008)
H13 (-C3)	0.164 (0.008)
H14 (-C4)	0.183 (0.009)
H15 (-C4)	0.177 (0.009)
H16 (-C6)	0.165 (0.009)
H17 (-C7)	0.154 (0.011)
H18 (-C7)	0.157 (0.011)
H19 (-C8)	0.161 (0.008)
H20 (-C8)	0.171 (0.008)
H21 (-C9)	0.156 (0.009)
H22 (-C9)	0.154 (0.009)
H23 (-C9)	0.155 (0.009)
H24 (-C10)	0.155 (0.011)
H25 (-C10)	0.156 (0.011)
H26 (-C10)	0.157 (0.011)

^a Atom types are displayed in Figure 1.

and $\mathbf{M}(t) \cong (1/2c) \sum_i \langle q_i \rangle \mathbf{r}_i(t) \times \mathbf{v}_i(t)$, where $\langle q_i \rangle$ represents the average partial charge of the i th atom. In Table 1, the atomic partial charges averaged over the 1 ns QM/MM MD trajectory as well as their standard deviations are given. The standard deviations of fluctuating atomic partial charges are fairly small, which is because the solute molecule, pinene, is relatively rigid and the surrounding nonpolar solvent molecules CCl_4 weakly perturb the electronic structure of the pinene.

Using the approximate $\mu(t)$ and $\mathbf{M}(t)$ with fixed atomic partial charges, we calculated the corresponding time-correlation functions and obtained the simulated IR absorption and VCD spectra (see the blue lines in Figures 3 and 4). The simulated C–C stretch, C–H bend, C=C stretch spectra with fixed charges are quantitatively similar to those with fluctuating charges. In contrast, the high-frequency C–H stretch, overtone, and combination spectra, as can be seen in Figures 3d and 4, are strongly dependent on whether the polarizable nature of the solute is taken into consideration or not. Particularly, if the solute is treated as a nonpolarizable molecule, the NIR absorption and VCD band intensities are negligibly smaller than those with fluctuating atomic partial charges. This is an interesting observation because it suggests that the polarizable nature of a given solute is quite important in the determination of the dipole and rotational strengths of high-frequency modes. On the other hand, the transition dipole and the rotational strengths of low-frequency modes are largely determined by structural fluctuation

amplitudes, i.e., standard deviations of fluctuating atomic coordinates and velocities. This can be understood by noting that the low-frequency modes are typically large-amplitude vibrations in comparison to those of high-frequency vibrations—note that the root-mean-square of vibrational displacement is inversely proportional to $(\mu\omega)^{1/2}$, where μ and ω are the reduced mass and angular frequency, respectively. Thus, the present computational studies clearly show that the polarizable nature of molecular vibrations is absolutely necessary to quantitatively describe NIR absorption and VCD phenomena of chiral molecules in condensed phases.

4. SUMMARY

In the present paper, we showed that the time-correlation function approaches with QM/MM MD simulation methods are useful to simulate not only mid-IR but also near-IR absorption and VCD spectra of chiral molecules in solutions. Not only the complicated solute–solvent interaction-induced effects on vibrational line shapes but also the polarization effects on dipole and rotational strengths were fully taken into account by employing QM treatment of the solute and MM description of interacting solvent molecules. Since the molecular vibrational dynamics is fully determined by quantum mechanically calculated potential energy surface with approximate QM/MM solute–solvent interactions, not only the fundamental but also the overtone and combination mode transitions were correctly included without relying on local mode approximations. However, due to the limited accuracies of classical expressions for the electric and magnetic dipole moments with approximate QM/MM solute–solvent interactions, the mid- and near-IR absorption and VCD spectra of low-frequency C–C stretching and C–H overtone modes were found to deviate from experimental results. We believe that such a lack of chemical accuracy in simulating NIR spectra can be overcome by employing a better QM method and by performing sufficiently long MD simulations. Nevertheless, we anticipate that the general strategy utilizing time-correlation function theory in combination with ab initio QM/MM MD simulation method will be of critical use in quantitatively simulating a variety of vibrational spectra including near-IR VCD.

AUTHOR INFORMATION

Corresponding Author

*E-mail: mcho@korea.ac.kr.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENT

This work was supported by the National Research Foundation of Korea (NRF) grants funded by the Korea government (MEST) (nos. 20090078897 and 20110020033) and by the financial support from the Korea Basic Science Institute (T31401) grant to M.C. Also, we thank an anonymous reviewer for suggesting us to replot Figures 3 and 4 with proper intensity scales.

REFERENCES

- (1) In *Near-Infrared Spectroscopy: Principles, Instruments, Applications*; Siesler, H. W., Ozaki, Y., Kawata, S., Heise, H. M., Eds.; Wiley-VCH: Weinheim, Germany, 2002.
- (2) Sage, M. L.; Jortner, J. *Adv. Chem. Phys.* **1981**, *47*, 293–322.
- (3) Child, M. S.; Halonen, L. *Adv. Chem. Phys.* **1985**, *57*, 1–58.
- (4) Donaldson, D. J.; Tuck, A. F.; Vaida, V. *Chem. Rev.* **2003**, *103*, 4717–4730.
- (5) Wolf, M.; Ferrari, M.; Quaresima, V. *J. Biomed. Opt.* **2007**, *12*, 062104.
- (6) Jobsis, F. F. *Science* **1977**, *198*, 1264–1267.
- (7) Chýlek, P.; Geldart, D. J. W. *Geophys. Res. Lett.* **1997**, *24*, 2015–2018.
- (8) Child, M. S. *Acc. Chem. Res.* **1985**, *18*, 45–50.
- (9) Henry, B. R. *Acc. Chem. Res.* **1987**, *20*, 429–435.
- (10) Abbate, S.; Castiglioni, E.; Gangemi, F.; Gangemi, R.; Longhi, G. *Chirality* **2009**, *21*, E242–E252.
- (11) Guo, C.; Shah, R. D.; Dukor, R. K.; Freedman, T. B.; Cao, X.; Nafie, L. A. *Vib. Spectrosc.* **2006**, *42*, 254–272.
- (12) Hsu, E. C.; Holzwarth, G. *J. Chem. Phys.* **1973**, *59*, 4678–4685.
- (13) Nafie, L. A.; Dukor, R. K.; Roy, J.-R.; Rilling, A.; Cao, X.; Buijs, H. *Appl. Spectrosc.* **2003**, *57*, 1245–1249.
- (14) Cao, X.; Shah, R. D.; Dukor, R. K.; Guo, C.; Freedman, T. B.; Nafie, L. A. *Appl. Spectrosc.* **2004**, *58*, 1057–1064.
- (15) Guo, C.; Shah, R. D.; Mills, J.; Dukor, R. K.; Cao, X.; Freedman, T. B.; Nafie, L. A. *Chirality* **2006**, *18*, 775–782.
- (16) Keiderling, T. A.; Stephens, P. J. *Chem. Phys. Lett.* **1976**, *41*, 46–48.
- (17) Abbate, S.; Longhi, G.; Givens, J. W.; Boiadjev, S. E.; Lightner, D. A.; Moscovitz, A. *Appl. Spectrosc.* **1996**, *50*, 642–643.
- (18) Abbate, S.; Longhi, G.; Boiadjev, S.; Lightner, D. A.; Bertucci, C.; Salvadori, P. *Enantiomer* **1998**, *3*, 337–347.
- (19) Castiglioni, E.; Lebon, F.; Longhi, G.; Abbate, S. *Enantiomer* **2002**, *7*, 161–173.
- (20) Abbate, S.; Longhi, G.; Santina, C. *Chirality* **2000**, *12*, 180–190.
- (21) Abbate, S.; Castiglioni, E.; Gangemi, F.; Gangemi, R.; Longhi, G.; Ruzziconi, R.; Spizzichino, S. *J. Phys. Chem. A* **2007**, *111*, 7031–7040.
- (22) Longhi, G.; Gangemi, R.; Lebon, F.; Castiglioni, E.; Abbate, S.; Pultz, V. M.; Lightner, D. A. *J. Phys. Chem. A* **2004**, *108*, 5338–5352.
- (23) Abbate, S.; Longhi, G.; Ricard, L.; Bertucci, C.; Rosini, C.; Salvadori, P.; Moscovitz, A. *J. Am. Chem. Soc.* **1989**, *111*, 836–840.
- (24) Bak, K. L.; Bludský, O.; Jørgensen, P. *J. Chem. Phys.* **1995**, *103*, 10548–10555.
- (25) Polavarapu, P. L. *Mol. Phys.* **1996**, *89*, 1503–1510.
- (26) Abbate, S.; Gangemi, R.; Longhi, G. *J. Chem. Phys.* **2002**, *117*, 7575–7586.
- (27) Gangemi, R.; Longhi, G.; Abbate, S. *Chirality* **2005**, *17*, 530–539.
- (28) Gangemi, F.; Gangemi, R.; Longhi, G.; Abbate, S. *Phys. Chem. Chem. Phys.* **2009**, *11*, 2683–2689.
- (29) Cheng, Y.-L.; Chen, H.-Y.; Takahashi, K. *J. Phys. Chem. A* **2011**, *115*, 5641–5653.
- (30) Lehmann, K. K. *J. Chem. Phys.* **1983**, *79*, 1098–1098.
- (31) Mills, I. M.; Robiette, A. G. *Mol. Phys.* **1985**, *56*, 743–765.
- (32) Ricard-Lespade, L.; Longhi, G.; Abbate, S. *Chem. Phys.* **1990**, *142*, 245–259.
- (33) Gangemi, F.; Gangemi, R.; Longhi, G.; Abbate, S. *Vib. Spectrosc.* **2009**, *50*, 257–267.
- (34) Marcott, C.; Faulkner, T. R.; Moscovitz, A.; Overend, J. J. *Am. Chem. Soc.* **1977**, *99*, 8169–8175.
- (35) Faulkner, T. R.; Marcott, C.; Moscovitz, A.; Overend, J. J. *Am. Chem. Soc.* **1977**, *99*, 8160–8168.
- (36) Gorbunov, R. D.; Nguyen, P. H.; Kobus, M.; Stock, G. *J. Chem. Phys.* **2007**, *126*, 054509.
- (37) Horníček, J.; Kaprálová, P.; P., B. *J. Chem. Phys.* **2007**, *127*, 084502.
- (38) Kinnaman, C. S.; Cremeens, M. E.; Romesberg, F. E.; Corcelli, S. A. *J. Am. Chem. Soc.* **2006**, *128*, 13334–13335.

- (39) Schmidt, J. R.; Corcelli, S. A.; Skinner, J. L. *J. Chem. Phys.* **2004**, *121*, 8887–8896.
- (40) Yang, S.; Cho, M. *J. Chem. Phys.* **2009**, *131*, 135102.
- (41) Yang, S.; Cho, M. *J. Chem. Phys.* **2005**, *123*, 134503.
- (42) Choi, J.-H.; Hahn, S.; Cho, M. *Biopolymers* **2006**, *83*, 519–536.
- (43) Choi, J.-H.; Hahn, S.; Cho, M. *Int. J. Quantum Chem.* **2005**, *104*, 616–634.
- (44) Choi, J.-H.; Lee, H.; Lee, K.-K.; Hahn, S.; Cho, M. *J. Chem. Phys.* **2007**, *126*, 045102.
- (45) Jeon, J.; Yang, S.; Choi, J.-H.; Cho, M. *Acc. Chem. Res.* **2009**, *42*, 1280–1289.
- (46) Bour, P.; Keiderling, T. A. *J. Phys. Chem. B* **2005**, *109*, 23687–23697.
- (47) Rhee, H.; June, Y.-G.; Lee, J.-S.; Lee, K.-K.; Ha, J.-H.; Kim, Z. H.; Jeon, S.-J.; Cho, M. *Nature* **2009**, *458*, 310–313.
- (48) Choi, J.-H.; Cheon, S.; Cho, M. *J. Chem. Phys.* **2010**, *132*, 074506.
- (49) Rhee, H.; Choi, J.-H.; Cho, M. *Acc. Chem. Res.* **2010**, *43*, 1527–1536.
- (50) Ringer, A. L.; MacKerell, A. D. *J. Phys. Chem. Lett.* **2011**, *2*, 553–556.
- (51) Brooks, B. R.; Brucoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* **1983**, *4*, 187–217.
- (52) Guest, M. F.; Bush, I. J.; Van Dam, H. J. J.; Sherwood, P.; Thomas, J. M. H.; Van Lenthe, J. H.; Havenith, R. W. A.; Kendrick, J. *Mol. Phys.* **2005**, *103*, 719–747.
- (53) Mulliken, R. S. *J. Chem. Phys.* **1955**, *23*, 1833–1840.
- (54) Abbate, S.; Longhi, G.; Kwon, K.; Moscovitz, A. *J. Chem. Phys.* **1998**, *108*, 50–62.
- (55) Scott, A. P.; Radom, L. *J. Phys. Chem.* **1996**, *100*, 16502–16513.

Accuracy of Calculated Chemical Shifts in Carbon 1s Ionization Energies from Single-Reference *ab Initio* Methods and Density Functional Theory

Alf Holme,[†] Knut J. Børve,^{*,†} Leif J. Sæthre,[†] and T. Darrah Thomas[‡]

[†]Department of Chemistry, University of Bergen, NO-5007 Bergen, Norway

[‡]Department of Chemistry, Oregon State University, Corvallis, Oregon 97331-4003, United States

 Supporting Information

ABSTRACT: A database of 77 adiabatic carbon 1s ionization energies has been prepared, covering linear and cyclic alkanes and alkenes, linear alkynes, and methyl- or fluoro-substituted benzenes. Individual entries are believed to carry uncertainties of less than 30 meV in ionization energies and less than 20 meV for shifts in ionization energies. The database provides an unprecedented opportunity for assessing the accuracy of theoretical schemes for computing inner-shell ionization energies and their corresponding chemical shifts. Chemical shifts in carbon 1s ionization energies have been computed for all molecules in the database using Hartree–Fock, Møller–Plesset (MP) many-body perturbation theory of order 2 and 3 as well as various approximations to full MP4, and the coupled-cluster approximation with single- and double-excitation operators (CCSD) and also including a perturbational estimate of the energy effect of triple-excitation operators (CCSD(T)). Moreover, a wide range of contemporary density functional theory (DFT) methods are also evaluated with respect to computing experimental shifts in C1s ionization energies. Whereas the top *ab initio* methods reproduce the observed shifts almost to within the experimental uncertainty, even the best-performing DFT approaches meet with twice the root-mean-squared error and thrice the maximum error compared to CCSD(T). However, a number of different density energy functionals still afford sufficient accuracy to become tools in the analysis of complex C1s photoelectron spectra.

1. INTRODUCTION

Experimental inner-shell ionization energies and electron structure theory are closely linked.¹ On the one hand, the ability to reproduce experimental energies may provide a stringent test for theoretical methods, for instance in the context of deriving new, improved energy density functionals. On the other hand, as experimentalists target increasingly complex molecules, theoretical ionization energies play an important role in assigning and interpreting photoelectron spectra. Using *m*-xylene as an example, the carbon 1s photoelectron spectrum consists of five closely spaced peaks originating from the five chemically inequivalent carbon atoms. Although it is possible to assign this spectrum unambiguously and hence obtain ionization energies corresponding to each of the unique atoms strictly on the basis of the recorded spectrum, one gains confidence in the results upon discovering that the experimentally derived ionization energy shifts are in good agreement with those predicted by theory.²

A more critical situation that demands reliable theoretical prediction arises with 1-pentyne, which can exist in two different conformers. In this case, the interest lies in determining the relative populations of the conformers. In this molecule, there are five inequivalent carbon atoms, with the possibility that the carbon atoms in one conformer have different ionization energies from those in the other. There are, thus, potentially 10 different peaks in the spectrum. In this case, it has been essential to have good theoretical predictions of relative ionization energies in order to obtain a reliable analysis of the spectrum.³

Many procedures have been proposed for estimating ionization energies and/or shifts in ionization energies: potential models, Koopmans' theorem, the equivalent-cores approximation, transition-state models, and hole-state calculations of various degrees of sophistication.^{4–6} In several previous studies,^{2,7–11} we found that useful results can be obtained with a Δ SCF approach, which involves separate self-consistent field calculations for initial and final states, with the effect from the singly ionized core on the valence electrons simulated by an effective core potential (ECP). The great advantage of this approach is that it is facilitated by the use of most *ab initio* or density functional theory (DFT) methods as implemented in standard electronic-structure codes, for the calculation of shifts in K-shell ionization energies for closed-shell molecules. The main disadvantage of the ECP approach is that it does not explicitly include the ionized and hence singly occupied core orbital. Similar to what is the case for the equivalent-cores approximation, this leads to an almost constant energy offset for the ionized state and precludes the calculation of absolute ionization energies. It does, however, allow for the calculation of chemical shifts, since the unknown energy offset very closely cancels in the subtraction. The errors associated with this simplified description are explored and quantified in this work by comparison to calculations that include an explicit, self-consistent core-hole orbital.

Received: September 20, 2011

Published: November 14, 2011

The Δ SCF/ECP approach implemented in terms of the hybrid density functional theory method B3LYP and polarized triple- ζ basis sets is reasonably economical in regard to computational resources; it can handle large molecules and has been found to give good predictions of relative carbon 1s ionization energies for a wide variety of compounds. Nonetheless, this approach has some specific failures. First, although the method reliably predicts the order of ionization energies for chemically inequivalent carbon atoms, it tends to overestimate the shifts by as much as 20%.^{2,7–10,12} Second, the method works well if we are comparing carbon atoms with the same hybridization but is less reliable if we compare the ionization energies of carbon atoms with different hybridization.

A number of comparative studies have been carried out in order to identify energy density functionals with superior performance in the calculation of ionization energies for a selected class of core states.^{4,13–18} Unfortunately, the comparison between theory and experiment is often complicated by rather small sets of experimental energies or data of either mixed, unknown, or low accuracy.⁶ To improve upon this situation, we have invested considerable effort into measuring carbon 1s ionization energies of small and medium-sized organic molecules as accurately as possible with present day electron spectrometers and third generation synchrotron light sources. The usefulness of this data set was demonstrated by Takahata and co-workers,¹⁹ who used our experimental energies to explore the accuracy of one particular choice of exchange (Perdew–Wang 1986) and correlation (Perdew–Wang 1991) functionals to the calculation of C1s ionization energies. As applied to shifts in 89 adiabatic C1s energies in F- and CH₃-substituted benzenes (relative to benzene), their favored approach gave an average absolute deviation (AAD) from the experimental values of 0.039 eV, which is marginally better than for the B3LYP/ECP approach described above (0.041 eV). The deviation was found to increase somewhat, to an AAD = 0.056 eV, as the chemical diversity was increased.¹⁹ This data set is still heavily biased however with 70% of the entries referring to F- and CH₃-substituted benzenes.

In order to provide a broader and more systematic basis for comparing theoretical methods with experimental results and with each other, we have compiled a database of 77 accurate adiabatic C1s ionization energies for 27 hydrocarbons and fluorohydrocarbons. Included are compounds containing carbon with sp, sp², and sp³ hybridization; aliphatic and aromatic compounds; and compounds with conjugated systems of double bonds. We have measured these experimental values with uniform accuracy at third generation synchrotron facilities. Most of these results have been previously published, but we include also 17 new experimental values, which have been measured using the same procedures that have been previously described. To serve the stated purpose, it is essential to establish the accuracy of the data set, as has been done in the course of the experiments.^{8,9,20} However, estimating the accuracy of an experiment is not easy, and in particular systematic errors may appear unnoticed. Here, the issue of experimental uncertainty is revisited by means of theory, working from the hypothesis that as the theory is systematically improved, the agreement with experimental results should approach the uncertainty of the experimental numbers.

The first part of this work is devoted to comparing experimental shifts in C1s ionization energies with calculated values obtained using a number of wave function-based *ab initio* methods of increasing sophistication, all in the spin-restricted

formalism. These methods include Hartree–Fock theory (HF), Møller–Plesset (MP) many-body perturbation theory of orders 2 and 3 as well as various approximations to full MP4, and coupled-cluster theory with the cluster operator containing single and double excitations (CCSD) and also including a perturbative estimate of the energy contribution from uncoupled triples (CCSD(T)). Apart from Hartree–Fock, these methods have in common that electron correlation is accounted for in terms of determinants that are excited relative to a single reference wave function, namely, the Hartree–Fock state. Formally, the computational cost of HF and B3LYP scales as the fourth power of the number of basis functions N , the corresponding scaling laws for the other methods being N^5 for MP2; N^6 for MP3, MP4D, MP4DQ, MP4SDQ and CCSD; and N^7 for CCSD(T). The methods that scale as the sixth power of N are listed in order of increasing proportionality factor, where the jump between MP3 and MP4SDQ represents a factor of about 3 and the subsequent jump to CCSD is much larger due to the iterative nature of the latter. CCSD(T) and several of the truncated MP4 approaches will be shown to provide highly reliable predictions of C1s shifts (within a few percent), essentially validating the experimental uncertainty estimates. Particular attention is paid to the presence of systematic departure from the general behavior when comparing carbon atoms of different hybridization.

Next, the present compilation of accurate core-level ionization energies will be used as a testing ground for energy density functionals. More specifically, 138 different energy functionals are assessed on their ability to reproduce shifts in C1s energies, covering all main classes of contemporary DFT functionals. An important part of this study is to compare 10 exchange and 10 correlation functionals, respectively, by systematically testing the 100 combination functionals to which they give rise. The comparison of experimental and calculated shifts confirms our previous experience that B3LYP tends to overestimate chemical shifts in C1s ionization energies. There are a number of better alternatives among the DFT functionals, one notable example being the exchange functional proposed by Becke and Roussel (BRx) combined with any of a number of different gradient-corrected correlation functionals.

2. COMPUTATIONAL

The Gaussian 03 set of programs²¹ was used with the B3LYP method to calculate optimized geometries, vibrational frequencies, and zero-point energies for the neutral molecules and the carbon-1s-ionized species. For these calculations, we used atom-centered Gaussian-type functions contracted to triple- ζ quality²² and augmented by polarization functions,²³ leading to C: [5s, 3p, 1d] and H: [3s, 1p]. We refer to this basis set as the TZP basis. For the core-ionized carbon atom, the corresponding nitrogen basis was used with all exponents scaled by a common factor of 0.9293, obtained by minimizing the energy of core-ionized methane.²⁴ The core of the ionized carbon atom was represented by the effective core potential (ECP) of Stevens et al.,²⁵ scaled to account for only one electron in the 1s shell.²⁶ Further details on these procedures can be found in refs 24 and 12. While relativistic effects contribute about 0.05 eV to the C1s ionization energy,²⁷ this converts into less than 0.04% in a chemical shift and is thus not considered further in this work.

With the optimized geometries determined as outlined above, single-point calculations were carried out using the HF, MP2,

Table 1. Experimental Adiabatic C1s Ionization Energies and Experimental and Theoretical (CCSD(T)/TZP) Shifts in C1s Ionization Energies, Relative to That of Methane (All Energies in eV)

molecule	atom	IE		shift		ref
		exptl	exptl	exptl	CCSD(T)	
methane	C1	290.689	0.000	0.000		20
ethane	C1	290.545	-0.144	-0.173		20
ethene	C1	290.695	0.006	0.024		20
ethyne	C1	291.128	0.439	0.451		20
propane	C1	290.345	-0.344	-0.332		24
	C2	290.473	-0.216	-0.250		
propene	C1	290.136	-0.553	-0.520		8
	C2	290.612	-0.077	-0.066		
	C3	290.671	-0.018	-0.019		
propyne	C1	290.226	-0.463	-0.443		39
	C2	290.778	0.089	0.117		
	C3	291.610	0.921	0.900		
butane	C1	290.266	-0.423	-0.395		24, 40
	C2	290.312	-0.377	-0.415		
1,3-butadiene	C1	290.060	-0.629	-0.607		8
	C2	290.683	-0.006	0.043		
1-butyne	C1	290.057	-0.632	-0.580		41
	C2	290.554	-0.135	-0.064		
	C3	291.410	0.721	0.702		
	C4	290.673	-0.016	0.003		
2-butyne	C1	291.291	0.602	0.574		41
	C2	290.012	-0.677	-0.671		
1-pentyne	C1	289.979	-0.710	-0.645		3
anti	C3	291.213	0.524	0.518		
1-pentyne	C1	290.017	-0.672	-0.641		3
gauche	C3	291.219	0.530	0.524		
2-pentyne	C1	291.229	0.540	0.541		41
	C2	289.919	-0.770	-0.777		
	C3	289.792	-0.897	-0.825		
	C4	291.120	0.431	0.406		
	C5	290.462	-0.227	-0.221		
1,3-pentadiene	C1	289.762	-0.927	-0.903		8
	C2	290.501	-0.188	-0.142		
	C3	290.247	-0.442	-0.408		
	C4	290.093	-0.596	-0.579		
	C5	290.565	-0.124	-0.114		
trans-3-hexene	C1	290.256	-0.433	-0.421		42
	C2	290.386	-0.303	-0.308		
	C3	289.879	-0.810	-0.745		
3-hexyne	C1	290.442	-0.247	-0.239		41
	C2	291.086	0.397	0.380		
	C3	289.726	-0.963	-0.909		
cyclohexane	C1	290.123	-0.566	-0.557		12
cyclohexene	C1	289.908	-0.781	-0.727		12
	C3	290.321	-0.368	-0.365		
	C4	290.257	-0.432	-0.427		
1,4-cyclo-	C1	290.086	-0.603	-0.559		12
hexadiene	C3	290.556	-0.133	-0.147		
benzene	C1	290.241	-0.448	-0.438		43

Table 1. Continued

molecule	atom	IE		shift		ref
		exptl	exptl	exptl	CCSD(T)	
toluene	C1	290.296	-0.393	-0.413		2
	C2	289.970	-0.719	-0.694		
	C3	290.101	-0.588	-0.585		
	C4	290.030	-0.659	-0.661		
	CH ₃	290.493	-0.196	-0.180		
<i>m</i> -xylene	C1	290.134	-0.555	-0.546		2
	C2	289.716	-0.973	-0.922		
	C4	289.779	-0.910	-0.896		
	C5	289.956	-0.733	-0.713		
	CH ₃	290.407	-0.282	-0.258		
<i>p</i> -xylene	C1	290.054	-0.635	-0.623		2
	C2	289.842	-0.847	-0.830		
	CH ₃	290.395	-0.294	-0.268		
fluorobenzene	C1	292.734	2.045	2.071		9
	C2	290.505	-0.184	-0.171		
	C3	290.592	-0.097	-0.068		
	C4	290.375	-0.314	-0.275		
1,3-difluoro-	C1	293.082	2.393	2.449		9
benzene	C2	290.781	0.092	0.107		
	C4	290.624	-0.065	-0.014		
	C5	290.914	0.225	0.293		
1,4-difluoro-	C1	292.867	2.178	2.226		9
benzene	C2	290.838	0.149	0.188		
4-fluorotoluene	C1	290.425	-0.264	-0.255		10
	C2	290.305	-0.384	-0.336		
	C3	290.363	-0.326	-0.326		
	C4	292.485	1.796	1.832		
	CH ₃	290.644	-0.045	-0.009		

MP3, MP4D, MP4DQ, MP4SDQ, CCSD, and CCSD(T)^{28–31} methods as well as an extensive set of density-functional methods as detailed in Figures 4 and 5, all in conjunction with the same basis sets as used during geometry optimization.

In a limited test of basis-set sensitivity in the computed core-level shifts, additional shift calculations were performed for all *ab initio* and DFT methods with the cc-pVTZ basis set³² for both ground and core-ionized states. The core of the ionized carbon atom was represented by the same ECP as in the B3LYP/TZP calculations, with the valence electrons described in the regular cc-pVTZ carbon set augmented by tight CV(2s,2p) functions.³³ Additionally, analogous calculations with the cc-pVDZ sets were carried out for all DFT functionals in the study and used together with cc-pVTZ in an extrapolation procedure toward the complete basis-set limit. For selected functionals, additional tests were performed with the large cc-pVQZ basis. In all cases, the relative ionization energies include zero-point-energy contributions as obtained from the vibrational frequencies calculated at the B3LYP/TZP level of theory.

To explore the accuracy of the ECP description of the ionized core, we compare results at the HF level of accuracy obtained with the ECP approximation and also with an explicit representation of the core hole. These calculations were performed using the DALTON code³⁴ and both TZP and cc-pVTZ basis sets.

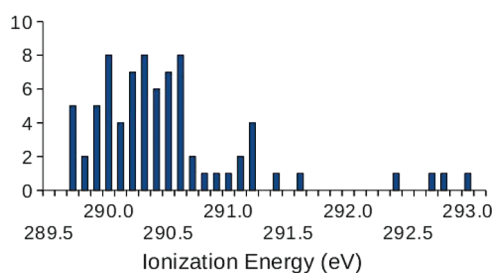


Figure 1. Distribution of experimental C1s ionization energies, in the database, cf. Table 1.

Table 2. Statistical Parameters Derived from Comparison of Experimental and Calculated Carbon 1s Ionization Energies Using *ab Initio* and DFT Methods Combined with the TZP Basis (Uncertainties in the Last Digit Indicated in Parentheses, Energies Given in meV)

method	ME ^b	RMSE ^c	MaxE ^d	slope ^{a,e}	intercept ^{a,e}	RMSD ^{a,f}
HF	-67	115	-309	0.885 (15)	37 (9)	62
MP2	124	140	305	1.040 (15)	-108 (7)	52
MP3	45	62	156	1.022 (11)	-36 (5)	40
MP4D	38	54	108	1.033 (10)	-27 (5)	35
MP4DQ	33	52	140	1.026 (9)	-23 (5)	33
MP4SDQ	8	29	98	1.007 (7)	-2 (3)	24
CCSD	12	33	95	1.015 (8)	-5 (4)	28
CCSD(T)	18	32	72	1.022 (7)	-11 (3)	24
B3LYP	-74	96	-265	0.926 (13)	46 (8)	52
G96B95	-69	95	-262	0.928 (14)	41 (8)	56
BRxP86	0	55	-214	0.928 (11)	-18 (6)	45
BRxKCIS	29	70	217	0.925 (14)	-44 (7)	53
SVWN	47	86	304	0.973 (19)	-56 (10)	72

^a Excluding carbon atoms directly bonded to fluorine. ^b Mean error; theoretical – experimental shift. ^c Root-mean-squared error. ^d Maximum error in any single datum. ^e Best-fit line, experimental vs theoretical shifts. ^f Root-mean-squared deviation between best-fit line and experimental shifts.

3. RESULTS AND DISCUSSION

3.1. The Database. Table 1 provides accurate measurements of 77 adiabatic C1s ionization energies for a wide range of hydrocarbons as well as four fluorine-containing organic molecules, spanning the range from 289.7 eV (in 3-hexyne) to 293.1 eV (in 1,3-difluorobenzene) and with a strong emphasis toward the low end of the interval, cf. Figure 1. The ionization energies are given with uncertainties of 30 meV, of which the larger part is associated with the internal calibration compound used, which in most cases is carbon dioxide. The relative ionization energies, i.e., chemical shifts, are believed^{8,9,20} to carry uncertainties of 10–20 meV, the reality of which is addressed in the present study by comparison with theory.

For cases such as ethyne where vibronic coupling of symmetry-related (localized) hole states leads to significant splitting and delocalization of the core–hole state, Table 1 lists the average of the core ionization energies originating from the indicated atoms. This is done in order to simplify a comparison to computed 1s ionization energies, which usually is conducted within the localized-hole formalism. The emphasis in this work lies on

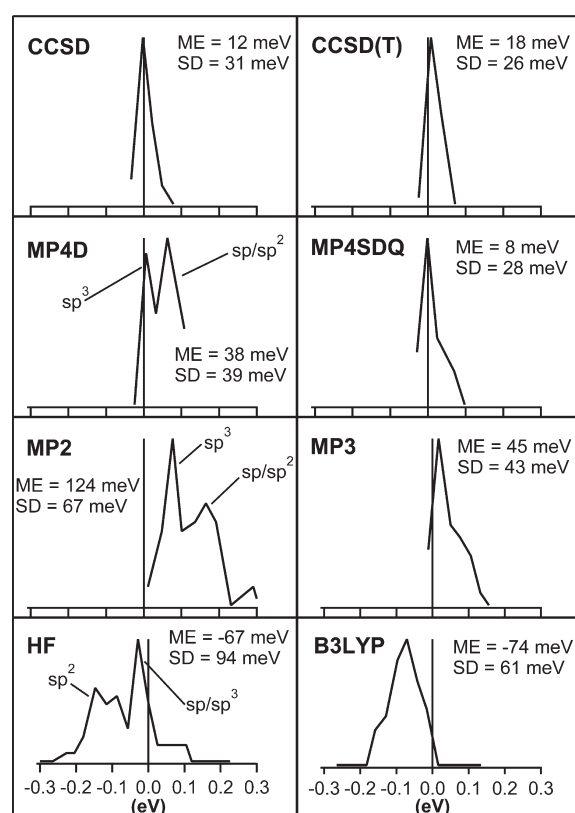


Figure 2. Error distribution in shifts in C1s ionization energies as computed with various electronic structure methods in conjunction with TZP basis sets. The mean error (ME) and standard deviation (SD) are included. Based on $N = 76$ energy shifts relative to methane.

chemical shifts in C1s ionization energies, and Table 1 shows also the shifts relative to methane as measured and as computed at the CCSD(T) level of theory. The corresponding data obtained for a wide range of other theoretical methods are provided as Supporting Information (SI).

3.2. Chemical C1s Shifts from *ab Initio* Methods and B3LYP. We will first compare the following single-reference *ab initio* models, HF, MP2, MP3, MP4D, MP4DQ, MP4SDQ, CCSD, and CCSD(T) with respect to how well they reproduce the experimental chemical shifts as measured by the mean error (ME), root-mean-squared error (RMSE), and the largest error (in magnitude) for any data point (MaxE), cf. top sections of Table 2, and with detailed statistical distributions in Figure 2. We also include B3LYP here, for comparison with previous work as well as a reference when making a wider screening of DFT functionals.

Referring to Table 2, HF and B3LYP give C1s shifts relative to methane that on the average are too low by about 0.07 eV (see also bottom panels of Figure 2). The error for the MP2 method is almost twice as large and of opposite sign, and MP2 comes across as a poor choice for computing chemical shifts in ionization energy. The mean error drops below 0.05 eV at the third order of many-body perturbation theory and is further reduced as fourth-order diagrams are added to the perturbation expansion. At the MP4SDQ, CCSD, and CCSD(T) levels of theory, the mean error is down to 0.01–0.02 eV. A similar development is found in the root-mean-squared error (RMSE), which levels off at 0.03 eV for the best *ab initio* methods. It is noteworthy that the more

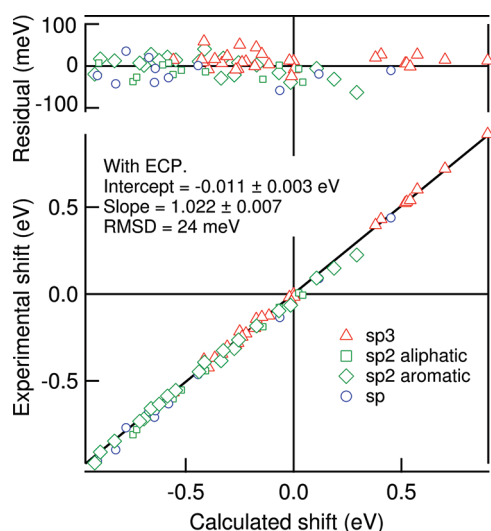


Figure 3. Experimental carbon 1s ionization-energy shifts (relative to methane) versus calculated values based on CCSD(T)/TZP with the core hole simulated by an effective core potential. The solid lines show least-squares fits to the data, and the graph labeled “Residual” shows the deviations of the points from the lines. Based on $N = 72$ data points (fluorocarbons excluded).

time-affordable SDQ approximation to full MP4 performs equally as well as the much more expensive coupled-cluster method CCSD(T).

The importance of triple excitations shows up in one statistical parameter: the largest error in any single shift value, MaxE. This number is 0.07 eV for CCSD(T) and 0.10 eV for CCSD and MP4SDQ and increasing in magnitude to +0.31 and -0.31 eV for MP2 and HF, respectively. Assuming that the errors are normally distributed around the mean, one may compute the standard deviation (SD) as $(\text{RMSE}^2 - \text{ME}^2)^{1/2}$. These values are quoted in Figure 2, to be discussed below. The maximum errors (MaxE) reported in Table 2 are consistent with errors that are normally distributed with the mean and standard deviation as stated, showing that neither the experimental values nor the computed ones contain outliers in the statistical sense.

In order to provide a more complete picture of the errors, we now discuss the error distributions for the different methods in Figure 2 in more detail. For the simpler methods, the error distribution is quite structured and indicates that there are systematic errors associated with subclasses of ionization sites. Closer scrutiny of the error distributions reveals that for HF, the peak near -0.05 eV is strongly dominated by sp - and sp^3 -hybridized carbon atoms, whereas errors between -0.1 and -0.2 eV represent sp^2 -hybridized carbon atoms exclusively. The error distribution is bimodal also in the case of MP2 and MP4D, but for these methods the large-error peak includes all sp and the vast majority of the sp^2 carbons, leaving sp^3 -hybridized carbon atoms for the low-error peak. The error distribution of B3LYP is similar to that of HF, albeit with a reduced difference between the two groups described for HF. The error distributions lose structure as the theoretical level is improved, and the distributions may be characterized compactly by the standard deviation (SD) about the mean error.

In Figure 3, we plot the shifts obtained experimentally for the 72 carbon atoms not directly bonded to fluorine, vs shift values calculated at the CCSD(T) level of theory. The four CF carbons,

i.e., carbon atoms directly bonded to fluorine, were omitted in this as well as subsequent analyses in order to avoid an undue bias caused by the four fluorinated carbon atoms at very high ionization energy, cf. Figure 1. The data are shown with symbols indicating the type of hybridization. A least-squares fit of a straight line to the data is indicated by the solid line, and the deviations of the points from the line (residuals) are shown in the upper part of the figure. As can be seen from the regression parameters included in Figure 3 and listed in Table 2, the line passes very close to the origin and has a slope close to 1 (1.022 ± 0.007).

Assuming that the linear fit is able to account for systematic errors in the computed chemical shifts, the remaining deviation represents an upper bound to the mean random error in the experimental shift values. In the case of CCSD(T), the root-mean-squared deviation (RMSD) of the points from the line is 24 meV, which is almost at the level of the expected error in the experimental shift values of 10–20 meV. Results for similar fits for all of the *ab initio* methods considered here are included in the right half of Table 2. CCSD and MP4SDQ group together with CCSD(T) in terms of low residual RMSD, a small intercept, and a slope very close to unity, with MP4D and MP4DQ trailing closely on all accounts, followed by MP3. The RMSD exceeds 0.05 eV for MP2, B3LYP, and HF, and while the intercept is particularly large for MP2, the slope is conspicuously off unity for HF and the hybrid DFT method. The latter statistics show that HF and B3LYP tend to overestimate the chemical shifts for hydrocarbons by about 10%.

3.3. Chemical C1s Shifts from Density Functional Theory.

The computational cost of advanced *ab initio* methods may be too high for routine calculation of chemical shifts for molecules with more than 10 non-hydrogen atoms, and it is clearly desirable to identify cheaper methods that can provide shifts with comparable accuracy. A large number of functionals have been proposed during the past 20 years, covering both hybrid and pure density functionals as well as combination functionals consisting of separate exchange and correlation functionals as well as integrated exchange-correlation functionals. To explore the accuracy in core-level shifts to be expected from different functionals, theoretical estimates of the 76 experimental C1s shifts accounted for above were obtained using a total of 138 different density functionals detailed as 100 combination functionals and an additional 30 hybrid functionals and eight pure functionals.

First, we consider the 100 energy density functionals obtained by combining 10 exchange functionals and 10 correlation functionals in all possible ways. The exchange functionals are denoted S (Slater), B (Becke1988), PW91 (Perdew–Wang1991), G96 (Gill1996), PBE (Perdew–Burke–Ernzerhof), PBEh (hybrid version of PBE), O (OPTX, by Handy–Cohan), BRx (Becke–Roussel1989), PKZB (Perdew–Kurth–Zupan–Blaha), and TPSS (Tao–Perdew–Staroverov–Scuseria), where simplified author lists are included within parentheses to provide a rationale for the acronyms. The acronyms are consistent with those in the Gaussian 09 user manual, which provides full references to the original publication of each functional. The corresponding list of correlation functionals is as follows: VWN (Vosko–Wilk–Nusair), PL (Perdew, local), LYP (Lee–Yang–Parr), P86 (Perdew, 1986), PW91 (Perdew–Wang, 1991), PBE (Perdew–Burke–Ernzerhof), B95 (Becke, 1995), PKZB (Perdew–Kurth–Zupan–Blaha), and KCIS (Krieger–Chen–Iafate–Savin).

In Figure 4, the root-mean-squared error (RMSE) as computed for the 100 combination functionals is presented in a heat

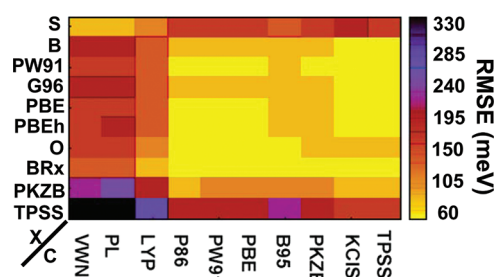


Figure 4. Root-mean-squared error (RMSE) in the C1s chemical shifts relative to methane predicted by selected exchange (X) and correlation (C) functionals and experimental results using the TZP basis. Based on 76 data points.

diagram, with exchange (correlation) functionals labeling rows (columns). It is immediately clear that the local correlation functionals (VWN and PL) lead to inferior results compared to the gradient-corrected ones, even though useful accuracy may be obtained in combination with the simple Slater exchange term, due to error cancellation between the local correlation and exchange functionals. Of the nonlocal correlation functionals, LYP stands out in the negative sense, with PKZB and B95 also performing worse than average. In the case of B95, the error is quite systematic, and a competitive RMSD value is obtained after applying a linear correction as described in the previous section. The difference in RMSE among the remaining functionals is quite small.

Turning to the exchange functionals labeling the rows in Figure 4, surprisingly large errors are found for PKZB and TPSS, irrespective of which correlation functional they are teamed up with. Apart from BRx, the other gradient-corrected exchange functionals group together with quite similar performance. BRx stands out with very small RMSE if combined with a gradient-corrected correlation functional different from LYP. This holds true also for the mean (signed) error, and for the root-mean-squared deviation (RMSD) from a best linear fit to the experimental data. According to Table 2, the lowest RMSE of 55 meV is obtained when BRx is combined with the P86 functional; this is comparable to our results for MP4D but much less accurate than the best *ab initio* methods. The RMSD value is slightly worse than for MP3: 45 vs 40 meV for 72 data points, as is the case for the largest error in any single data point (MaxE), -0.21 eV for BRxP86 and $+0.16$ eV for MP3. This is less accurate than hoped for, and we extended the study by 38 additional functionals of various origins, including the M06 functionals by Truhlar and co-workers (M06, M06HF, M062X, and M06L) and a number of hybrid functionals (B3LYP, B3P86, B3PW91, B1LYP, B1B95, O3LYP, X3LYP, mPW1PW91, mPW1LYP, mPW1PBE, mPW3-PBE, B98, B971, B972, PBE1PBE, BHandH, BHandHLYP, tHCTHhyb, HSEh1PBE, HSE2PBE, PBEh1PBE, ω B97XD, ω B97, ω B97X, TPSSh, LC- ω PBE, and CAM-B3LYP) as well as seven additional pure functionals (VSXC, HCTH, HCTH93, HCTH147, HCTH407, tHCTH, and B97D).

The root-mean-squared errors (RMSE) for the additional functionals are presented in Figure 5 along with data for five of the combination functionals to facilitate a comparison. Seventeen of the 43 functionals included in the figure have very similar RMSE values below 65 meV, i.e., at level with or better than MP3, but none is improving on the BRxP86 functional. The RMSE increases steadily through the next 12 functionals, before leveling out just below 100 meV.

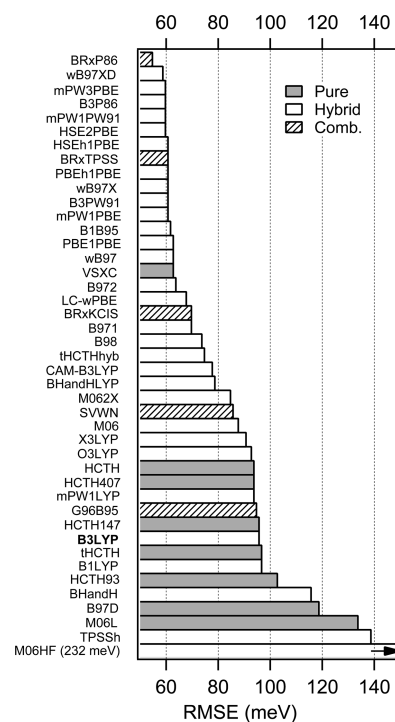


Figure 5. Root-mean-squared error (RMSE) in chemical shifts predicted by pure and hybrid functionals and experimental results using the TZP basis. Five combination functionals are included for comparison. Based on 76 data points.

In general, it is difficult to spot any correlation between performance and how the functionals were constructed, except that PBE and LYP are clearly over- and under-represented, respectively, in the high-performance group of functionals. A small class of functionals shows clearly inferior performance, led by M06L, M06HF, and the hybrid TPSSh functional. B3LYP, which is highlighted in the figure, is performing below average and for instance shows a higher RMSE than the local SVWN functional. Admittedly, the error is more systematic for B3LYP than for SVWN, as evident from the significantly lower RMSD after applying a linear-fit correction, cf. the rightmost column in Table 2.

A recent study by Takahata et al. shows that the PW86x-PW91c functional is able to predict carbon 1s ionization energies with high accuracy,¹⁹ as indicated by an average absolute deviation of 45 meV and a maximum absolute deviation of 130 meV from experimental values. The largest error was actually found for methane (130 meV) and with substantial negative errors for some of the multiply fluorinated benzenes and 1,3-pentadiene. In effect, a fairly large maximum error of 236 meV arises in their computed chemical shift for C1 in 1,3-pentadiene, using methane as the reference compound. For comparison, the corresponding numbers for the BRxP86 functional in conjunction with TZP bases are 36 meV (AAD) and 214 meV (MaxE). In fact, whereas more than 20 of the DFT functionals considered here display average absolute deviations of 45 meV or better, the maximum absolute deviation does not drop below 200 meV. The largest errors in our study were found for 1,3-difluorobenzene, 1,3-pentadiene, and 1-butyne, of which the latter was not included in ref 19.

3.4. Core-Hole Description and Basis Set Effects. It is thus clear that, even the best of the theoretical models is unable to

Table 3. Results of a Linear-Fit Comparison of C1s Shifts Computed at the Hartree–Fock/TZP Level of Theory, Using an Explicit Core–Hole Model and an Effective Core Potential (ECP) Model for the Ionized Core

data set	slope, s^a	intercept, μ (meV) ^a	RMSE (meV)
all	0.999	−16	24
sp ³	1.009	−8	11
sp ²	0.998	−9	14
aliphatic	0.983	−22	15
aromatic	1.004	−4	14
sp	0.984	−59	53

^a core hole shift = $s \times (\text{ECP shift}) + \mu$.

predict the shifts in C1s ionization energy better than to within an RMSE of about 30 meV, which is approaching but still larger than the estimated uncertainty in the experimental results of 10 to 20 meV. It is, therefore, of interest to explore whether the accuracy is limited by the treatment of electron correlation, or possibly by approximations common to the various methods such as the molecular geometries or the use of an effective core potential to describe the ionized core. A partial answer to this question comes from the statistical data already presented. The root-mean-squared error (RMSE) listed in Table 2 is clearly lowest for the high-level methods, MP4SDQ–CCSD(T), largely due to a corresponding reduction in mean error, i.e., a method-dependent uniform shift, as is evident from Figure 2. The standard deviation (SD) about the mean error is seen to be about 30 meV for the high-end methods, with a rapidly increasing SD value as the correlation treatment becomes less complete. However, a standard deviation of 30 meV is still 50% higher than the estimated experimental uncertainty, and the latter is then either too optimistic or there is a source of random error common to the (high-end) computational methods. One possible candidate for such an error is the approximate scheme used to describe the ionized atomic core, to be explored next.

3.4.1. ECP vs Explicit Core Hole. All of the calculations described above used a one-electron effective core potential (ECP) to simulate the effects of the ionized core on the valence electrons. To investigate the possible errors arising from this approximation, all of the chemical shifts in the database were recomputed at the Hartree–Fock level of theory using an explicit and self-consistent core-hole description. These results will now be compared with a similar set of calculations using the ECP, keeping molecular geometries and the basis set the same. A convenient way to analyze these results is by relating the core-hole-based shifts to those obtained with an ECP through a linear regression model, cf. Table 3. Importantly, the two sets of chemical shifts agree closely, as is evident from a slope of 0.9985 and an intercept of −16 meV for the correlation line, meaning that the shift data computed by means of the ECP lie 16 meV higher than those obtained using an explicit core hole. While this average picture applies equally well to sp³ and aromatic sp² carbons, larger errors occur for carbons with triple bonds and to some extent also to aliphatic sp² carbons. For sp-hybridized carbon atoms, the intercept triples to −59 meV and the slope drops to 0.984.

The observation that the ECP model faces limitations in the case of triply bonded carbon atoms is not completely unexpected, as the short C–C distance in the acetylenic moiety should induce stronger orbital relaxation effects in the core orbital of sp carbons

than other carbons. On the other hand, the short C–C distance also implies a stronger tendency to delocalize the core hole over the triply bonded pair of atoms, which is not at all well described at the self-consistent-field level of theory. Hence, an ECP description may still offer the best choice, and the question is whether it is possible to improve upon this model by simple means. One step in this direction is to assume that the change to the C1s shift due to inclusion of an explicit core hole is transferable between electronic-structure methods. If true, this implies that the chemical shift obtained with a given method may be corrected according to

$$\Delta\text{IE}(\text{M, corrected}) = \Delta\text{IE}(\text{M, ECP}) + \Delta\text{IE}(\text{HF, corehole}) - \Delta\text{IE}(\text{HF, ECP}) \quad (1)$$

where “M” refers to any electronic-structure method such as MP2 or CCSD(T). However, if this correction is applied to our CCSD(T)/TZP data, the root-mean-squared error (RMSE) drops only slightly, from 32 to 26 meV. The standard deviation about the mean error remains the same within 1 meV, as is the case also for the root-mean-squared deviation (RMSD) about a best-fit line. Hence, it appears that the use of an effective potential to represent the ionized core induces a very small systematic error in the computed C1s shifts, primarily associated with sp carbons. Correcting by reference to proper hole-state calculations appears relevant only for the most accurate of *ab initio* methods. The possible coupling between the ECP–core-hole correction and choice of basis set is commented upon below.

3.4.2. Basis Set Effects. The default basis set used in the calculations reported thus far is of triple- ζ plus polarization quality and based on the Dunning–Huzinaga set. To explore the impact of the basis set on the quality of the theoretical chemical shifts, the full set of calculations was repeated using the same methods and molecules as previously but now employing the cc-pVTZ basis. This basis set is better than TZP in terms of the polarization space and total number of basis functions, but it is fully contracted in the core orbital space. For this reason, two sets of tight s and p functions were added to the carbon atom undergoing core ionization. We note that the use of this basis set was advocated by Cavagliasso and Chong,¹³ with the slight difference that they apply the addition of compact core functions to all non-hydrogen atoms and also in the initial state. We have confirmed that this difference has negligible impact on the computed relative ionization energies.

The new C1s shifts obtained with the cc-pVTZ basis were compared with the experimental shifts, as reported in Table 4, and the RMSE values of selected methods are compared for the two basis sets. Except for MP2, the introduction of cc-pVTZ leads to poorer agreement with experimental results for all methods. For MP4SDQ and CCSD(T), the RMSE values increase by about 10 meV, while much larger changes are found for most of the DFT methods as well as HF. Quite uniformly for the DFT methods, the mean error is shifted toward lower values by about 50 meV upon changing from TZP to cc-pVTZ. The deviation from a linear best fit to experimental shift values also deteriorates with cc-pVTZ, resulting in RMSD values that typically increase by 20–30 meV. Exceptions to this trend are primarily functionals that represent exchange interaction only through the Slater term; they systematically give too high C1s shifts and hence benefit from the general reduction in shift values observed for cc-pVTZ. Moreover, functionals that include the KCIS correlation functional display less basis-set sensitivity, as illustrated by the representative examples in Table 4.

Table 4. Basis-Set Dependency of the Agreement between Experimental and Theoretical Shifts in Carbon 1s Ionization Energies As Measured by the Root-Mean-Squared Error (RMSE, in meV, based on 77 Data Points)^a

method	basis set					
	TZP	cc-pVTZ	cc-pVDZ	cc-pVQZ	X3(D,T)	X3(T,Q)
HF	115	151				
MP2	140	117				
MP4SDQ	29	39				
CCSD(T)	32	40				
B3LYP	96	149	262	139	105	132
G96B95	95	139	228		108	
BRxP86	55	98	188		74	
BRxKCIS	70	85	177	72	72	66
SVWN	86	91	167		72	

^aIn addition to the TZP and cc-pVnZ (n = D, T, and Q) basis sets, the leftmost part of the table reports on energies obtained with a two-point extrapolation method X3(n,m), where n and m are ordinals for cc-pVnZ bases.

The inferior results obtained with the cc-pVTZ basis beg for an explanation. The TZP basis is state adapted in the sense that different orbital exponents are used for the neutral and core-ionized carbon atom.³⁵ In more detail, the basis set used for the core-ionized carbon atom starts from the corresponding nitrogen basis, but with the orbital exponents uniformly scaled by a factor chosen to minimize the energy of core-ionized methane.²⁴ Hence, the TZP basis tends to give a low C1s energy for methane and presumably an energy penalty that increases with departure from a methane-like core. Possibly even more important for TZP outperforming cc-pVTZ is the fact that the scaling factor used for the ionized-carbon basis was optimized in the presence of the effective core potential.

Table 4 also reports on shift data obtained with selected DFT methods in conjunction with the double(D)- and quadruple(Q)- ζ bases cc-pVnZ, where n = 2, 3, and 4 corresponds to D, T, and Q, as well as those obtained in a single-parameter extrapolation based on two consecutive members of the correlation-consistent basis sets. The infinite-basis limit is estimated as³⁶ $E_\infty = ((1 + n^{-1})^3 E_{n+1} - E_n) / ((1 + n^{-1})^3 - 1)^{-1}$. Employment of the D and T terms in this extrapolation scheme produces shift data that approach the accuracy of those obtained with the TZP basis. For the BRxKCIS functional, extrapolation based on the T and Q data produces chemical C1s shifts that are marginally better than those computed with TZP. Hence, the TZP basis appears to be a very efficient tool for obtaining accurate shift estimates.

3.5. Corrected Shifts. The results discussed above show that if used with TZP bases, MP4DQ, MP4SDQ, CCSD, and CCSD(T) all provide chemical shifts in C1s ionization energies with a root-mean-squared error (RMSE) of about 30 meV or less for a mixed-hybridization set of atoms. Inclusion of a core-hole–ECP correction term does not change this picture. The less costly of the computational methods examined here, Hartree–Fock and density functional theory, lead to RMSEs from 55 meV in the best case (BRxP86) to more than 300 meV in the most unfavorable case, cf. Figure 4. It would be useful to be able to obtain reliable predictions of ionization energies from lower-level calculations. One approach is to remove a large part of the systematic errors in the computed numbers by invoking a linear correction, calibrated toward

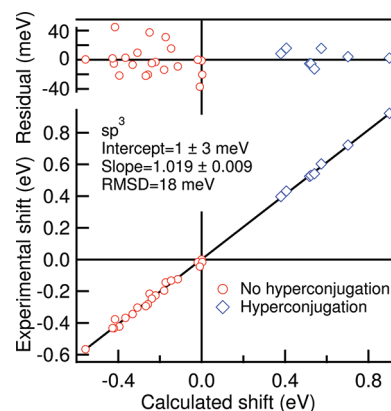


Figure 6. Shifts in carbon 1s ionization energies relative to that of methane for sp^3 -hybridized carbon atoms. Horizontal axis: Calculated shifts using the CCSD(T)/TZP level of theory. Vertical axis: Experimental shifts. The solid line shows a least-squares fit to the data.

experimental shift data. This implies the construction of semi-empirical predictions to experimental shifts, by way of $\Delta I E_{\text{exp,pred}} = m \times \Delta I E_{\text{theo}} + b$, where regression parameters are exemplified in Table 2 together with root-mean-squared deviation (RMSD) between experimental and corrected-computational shift data. The RMSD from the fit line is down to half of the RMSE for HF, B3LYP, and MP2, with significant improvements also for the other methods. The lowest RMSD value among the DFT methods (41 meV, cf. SI) is obtained for BRx in combination with LYP or either of the two local correlation functionals included in the study (VWN, PL), thus competing well with MP3 but still clearly inferior to the truly high-end *ab initio* methods.

Still higher accuracy is possible at the cost of losing generality if the linear correction is calibrated for subsets of carbon atoms of specific hybridization. In the next section, we consider this approach for CCSD(T), in order to examine the estimated uncertainty in the experimental ionization energies of 10–20 meV.

3.6. A Systematic View on the Basis of Hybridization. That predicting correctly the changes of ionization energies with changes in hybridization represents a challenge to computational models can be seen clearly in the series ethane, ethene, and ethyne, for which the carbon 1s ionization energy increases monotonically. The CCSD(T)/TZP model overestimates this increase by 5% and underestimates it by 3% if the ECP–hole-state correction mentioned in section 2 is included. Previous HF calculations³⁷ show that the predicted shift between ethane and ethene is sensitive to the choice of basis set and that the sign of the shift can be either positive or negative, depending on the basis set. The difference, ethene – ethane, arises from a cancellation of an initial-state effect of 0.32 eV^{5,37} by a comparable relaxation energy in the final state, and the theoretically calculated ionization energy is, therefore, sensitive to the details of the calculation.

The preceding paragraph indicates that even our most accurate computational approach, CCSD(T)/TZP, suffers from systematic errors in computed C1s energy shifts involving carbon atoms of different hybridization. To bypass this problem in order to establish upper bounds to the unsystematic error in the experimental shift values, we explore in some detail the agreement between computed and experimental values for subsets of atoms that share the formal hybridization.

A comparison of experimental and calculated ionization energies for 29 sp^3 hybridized carbon atoms is shown in Figure 6 for

Table 5. Comparison of Experimental Shifts^a in C1s Ionization Energies to Those Computed with CCSD(T) in Combination with TZP and cc-pVTZ Bases, Respectively (Uncertainties in the Last Digit Indicated in Parentheses, Energies Given in meV)

	ME ^b	RMSE ^c	slope	intercept	RMSD ^d
TZP basis					
all hybridizations	18	32	1.022 (7)	-11 (3)	24
sp ³ all (N = 28) ^e	-1	20	1.019 (9)	1 (3)	18
CH ₃ (N = 16) ^e	7	21	1.030 (11)	-6 (4)	16
CH ₂ (N = 12) ^e	-12	19	1.007 (10)	12 (4)	15
sp ² all (N = 33) ^{e,f}	26	33	0.989 (11)	-31 (6)	21
aliphatic (N = 12) ^e	35	38	1.011 (20)	-30 (9)	17
aromatic (N = 21) ^{e,f}	21	30	0.978 (13)	-30 (7)	20
sp all (N = 11) ^e	37	46	1.011 (21)	-32 (13)	29
cc-pVTZ basis					
all hybridizations	4	40	1.021 (11)	0 (6)	40
sp ³ all (N = 28)	-3	16	1.021 (1)	4 (3)	14
sp ² all (N = 33)	-10	35	1.034 (2)	25 (10)	33
sp all (N = 11)	76	81	0.984 (2)	-83 (12)	28

^aME and RMSE are based on 76 shift values, whereas slope, intercept, and RMSD are based on 72 shifts (fluorocarbons excluded). ^bMean error in the theoretical shifts. ^cRMSE between theoretical and experimental shifts. ^dRMSD from fit line. ^eThe number of data points used in the regression. ^fAn additional four data points representing CF carbons were included in ME and RMSE.

the CCSD(T) calculations. The solid line shows a least-squares fit of a straight line to the data, and the parameters of this line are listed in the second line of numbers in Table 5. We see that the slope and intercept are nearly the same as for the overall comparison, but that the root-mean-squared deviation (RMSD) is much smaller. From the residuals, plotted in the upper portion of Figure 6, we see that there is no systematic deviation of the points from the line. The root-mean-squared difference (RMSE) between experiment and theory is 20 meV compared with 32 meV for the overall comparison, and the RMS difference from the trend line is only 18 meV compared with 24 meV for the overall comparison. These differences, 20 and 18 meV, are comparable to the estimated uncertainties in the experimental data, indicating that the CCSD(T) calculations are describing these shifts within the experimental uncertainties. This is a strong indication that, at least for sp³ carbons, the estimated uncertainty in the experimental data, of 10–20 meV, is correct.

The points on the left-hand side of Figure 6 represent sp³-hybridized carbon atoms that are not significantly influenced by hyperconjugation—primarily those in fully saturated alkanes but also carbon atoms in alkenes and alkynes at positions where hyperconjugation does not have an influence. By contrast, those on the right are strongly influenced by hyperconjugation of the type found in propyne, for instance. In propyne, there is significant electron transfer from C3 (the carbon atom with sp³ hybridization) to the C1 carbon via contributions from the mesomeric form HC⁻=C=CH₃⁺. Figure 6 shows that the theoretical calculations provide satisfactory predictions of the effect of hyperconjugation on these ionization energies.

Table 5 also gives regression parameters derived for each hybridization separately and also selected subsets within each hybridization, by comparing the experimental shifts with those

predicted using CCSD(T)/TZP in the same manner as just discussed in detail for sp³. Focusing on sp² carbons, we find that the root-mean-squared error is noticeably greater than that of sp³ carbons, and that this may be traced to a corresponding difference in mean (signed) error. Accordingly, the intercept of the correlation line is significant while the slope remains within a couple of percent from 1. Importantly, the RMSD from the regression line is not much larger than the estimated experimental uncertainties, 21 meV for CCSD(T), and we conclude that an estimated uncertainty of 10–20 meV appears realistic.

For sp hybridization, the errors in the CCSD(T) shifts are somewhat larger than for the other hybridizations. This is evident from an RMSE of 46 meV compared to 33 and 20 meV for the sp² and sp³, respectively. After applying a linear regression model as reported in Table 5, the root-mean-squared deviation from the fit line to experimental shifts is about 30 meV. This number is significantly higher than the estimated uncertainty of 10–20 meV in the experimental energies. From the discussion in section 3.4.1, one realizes that part of this larger error may be ascribed to deficiencies in the effective-potential model (ECP) and possibly also the localized-hole model. Hence, while we are not able to demonstrate that the experimental energies for sp carbons are as accurate as for the other hybridizations, we are not in a position to prove otherwise.

Also included in Table 5 are statistics obtained from computed energies obtained with CCSD(T) in conjunction with cc-pVTZ bases. The RMSE values are similar to those obtained with the TZP basis, except for a much larger mean error and hence larger RMSE for the sp carbons. While this may be corrected for in a linear regression model, the models for the different hybridizations are sufficiently different to give a high overall RMSD of 40 meV for cc-pVTZ, compared to 24 meV for the TZP basis.

4. CONCLUDING REMARKS

A database of 77 adiabatic carbon 1s ionization energies has been prepared, covering linear and cyclic alkanes and alkenes, linear alkynes, and methyl- and fluoro-substituted benzenes. Individual entries are given with uncertainties of less than 30 meV in ionization energies. Comparison to theoretically computed chemical shifts in C1s energies confirms that the uncertainty in the experimental shift values is in the range 10–20 meV, at least for sp³ and sp² carbons. From the perspective of developing and testing new energy density functionals, the database constitutes a potentially useful supplement to existing compilations of accurate experimental data suitable for quantum chemical modeling. For instance, even the GMTKN30 superdatabase,³⁸ which includes data from thermochemistry, molecular structures, reaction kinetics, physical properties, and also proton affinities and valence ionization energies, does not contain core-level ionization energies. While it may be argued that the calculation of core-level energies is a specialized branch of computational chemistry, introduction of the ECP model facilitates the calculation of shifts in K-shell ionization energies on a routine basis.

The comparisons presented here show that the high-end *ab initio* electronic-structure methods MP4SDQ, CCSD, and CCSD(T), when used with scaled, polarized valence triple- ζ bases and an effective potential for the ionized core, are able to predict chemical shifts in carbon 1s ionization energies within a few percent and with an RMSE of about 30 meV. This is for a set

of carbon atoms that include all common hybridizations. There are systematic differences in the ability to reproduce shifts among carbons of sp^3 , sp^2 , and sp hybridizations. The mean and root-mean-squared errors are larger for sp carbons, although some of this discrepancy may be removed if we make a correction for the differences between calculations based on the ECP and those done with an explicit core hole. In general, the use of an effective ionized-core potential provides a very good approximation to the proper hole-state calculation.

Although CCSD(T) stands out as particularly robust and accurate as judged from the smallest worst-error (MaxE), MP4SDQ scales much more favorably in terms of computer time and still has almost as good predictive power with respect to shifts in carbon 1s ionization energies. Hartree–Fock and the hybrid-density-functional method B3LYP significantly overestimate C1s shifts within each single class of carbons that share hybridization as well as across the board, by as much as 15%. Due to the systematic nature of the errors, a linear fit of the computed shifts to the experimental ones provides regressional parameters that make it possible to correct the less costly HF, B3LYP, and MP2 calculations to give predictions of shifts that are approaching the experimental uncertainties within each hybridization. For most of the *ab initio* methods, the calculated shifts are in closer agreement with the experimental shifts using the scaled TZP basis set, as compared to the cc-pVTZ basis. This is convenient since the TZP basis is the smaller and less computationally demanding of the two.

After testing 138 contemporary density-functional methods, none is found to compete with the high-end wave-function-based methods with respect to prediction power. The BRx exchange functional in combination with a gradient-corrected correlation functional does represent a cost-efficient alternative if a root-mean-squared error in excess of 50 meV is acceptable, and this is still better than MP3 and similar to MP4DQ. However, the largest error in any single shift datum is consistently found to be significantly larger for the DFT methods, with none getting below 200 meV among the 76 shift values considered here.

■ ASSOCIATED CONTENT

S Supporting Information. Basis sets and molecular geometries for the neutral as well as each core-ionized state are provided as optimized at the B3LYP/TZP level of theory. Electronic and zero-point energies for all neutral and core-ionized states, for all *ab initio* and DFT methods that are considered in the article, using TZP and cc-pVTZ basis sets and, in the case of DFT, also cc-pVDZ and cc-pVQZ (for selected functionals). At the Hartree–Fock level, ionized-state energies are provided on the basis of an explicit core hole as well as an effective-core-potential (ECP) description, while for all other methods, ionized-state energies are obtained with an ECP model. This material is available free of charge via the Internet at <http://pubs.acs.org/>.

■ AUTHOR INFORMATION

Corresponding Author

*Phone: +47 5558 3365. Fax: +47 5558 9490. E-mail: knut.borve@kj.uib.no

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENT

A.H., L.J.S., and K.J.B. thank the Nordic Research Board (NORDFORSK), the Norwegian High Performance Computing Consortium NOTUR, and the EC Transnational Access to Research Infrastructure Program (TARI) for support.

■ REFERENCES

- (1) Fadley, C. S. *J. Electron Spectrosc. Relat. Phenom.* **2010**, *178–179*, 2–32.
- (2) Myrseth, V.; Sæthre, L. J.; Børve, K. J.; Thomas, T. D. *J. Org. Chem.* **2007**, *72*, 5715–5723.
- (3) Holme, A.; Sæthre, L. J.; Børve, K. J.; Thomas, T. D. *J. Mol. Struct.* **2009**, *920*, 387–392.
- (4) Triguero, L.; Plashkevych, O.; Pettersson, L. G. M.; Ågren, H. *J. Electron Spectrosc. Relat. Phenom.* **1999**, *104*, 195–207.
- (5) Børve, K. J.; Thomas, T. D. *J. Electron Spectrosc. Relat. Phenom.* **2000**, *107*, 155–161.
- (6) Shim, J.; Klobukowski, M.; Barysz, M.; Leszczynski, J. *Phys. Chem. Chem. Phys.* **2011**, *13*, 5703–5711.
- (7) Thomas, T. D.; Sæthre, L. J.; Børve, K. J.; Bozek, J. D.; Huttula, M.; Kukk, E. *J. Phys. Chem. A* **2004**, *108*, 4983–4990.
- (8) Thomas, T. D.; Sæthre, L. J.; Børve, K. J.; Gundersen, M.; Kukk, E. *J. Phys. Chem. A* **2005**, *109*, 5085–5092.
- (9) Carroll, T. X.; Thomas, T. D.; Bergersen, H.; Børve, K. J.; Sæthre, L. J. *J. Org. Chem.* **2007**, *71*, 1961–1968.
- (10) Carroll, T. X.; Thomas, T. D.; Sæthre, L. J.; Børve, K. J. *J. Phys. Chem. A* **2009**, *113*, 3481–3490.
- (11) Sæthre, L. J.; Børve, K. J.; Thomas, T. D. *J. Electron Spectrosc. Relat. Phenom.* **2011**, *183*, 2–9.
- (12) Oltedal, V. M.; Børve, K. J.; Sæthre, L. J.; Thomas, T. D.; Bozek, J. D.; Kukk, E. *Phys. Chem. Chem. Phys.* **2004**, *6*, 4254–4259.
- (13) Cavigliasso, G.; Chong, D. P. *J. Chem. Phys.* **1999**, *111*, 9485–9492.
- (14) Chong, D. P.; Aplincourt, P.; Bureau, C. *J. Phys. Chem. A* **2002**, *106*, 356–362.
- (15) Takahata, Y.; Chong, D. P. *J. Electron Spectrosc. Relat. Phenom.* **2003**, *133*, 69–76.
- (16) Takahashi, O.; Pettersson, L. G. M. *J. Chem. Phys.* **2004**, *121*, 10339–10345.
- (17) Chong, D. P. *J. Electron Spectrosc. Relat. Phenom.* **2005**, *148*, 115–121.
- (18) Segala, M.; Chong, D. P. *J. Electron Spectrosc. Relat. Phenom.* **2010**, *182*, 141–144.
- (19) Takahata, Y.; Marques, A. D.; Custodio, R. *THEOCHEM* **2010**, *959*, 106–112.
- (20) Myrseth, V.; Bozek, J. D.; Kukk, E.; Sæthre, L. J.; Thomas, T. D. *J. Electron Spectrosc. Relat. Phenom.* **2002**, *122*, 57–63.
- (21) 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 C.02; Gaussian, Inc.: Wallingford, CT, 2004.
- (22) Dunning, T. H., Jr. *J. Chem. Phys.* **1971**, *55*, 716–723.
- (23) Raghavachari, K.; Binkley, J. S.; Seeger, R.; Pople, J. A. *J. Chem. Phys.* **1980**, *72*, 650–654.

- (24) Karlsen, T.; Børve, K. J.; Sæthre, L. J.; Wiesner, K.; Bässler, M.; Svensson, S. *J. Am. Chem. Soc.* **2002**, *124*, 7866–7873.
- (25) Stevens, W. J.; Basch, H.; Krauss, M. *J. Chem. Phys.* **1984**, *81*, 6026–6033.
- (26) Karlsen, T.; Børve, K. J. *J. Chem. Phys.* **2000**, *112*, 7979–7985.
- (27) Chong, D. P. *J. Chem. Phys.* **1995**, *103*, 1842–1845.
- (28) Čížek, J. *Adv. Chem. Phys.* **1969**, *14*, 35–89.
- (29) Purvis, G. D.; Bartlett, R. J. *J. Chem. Phys.* **1982**, *76*, 1910–1918.
- (30) Scuseria, G. E.; Janssen, C. L.; Schaefer, H. F., III. *J. Chem. Phys.* **1988**, *89*, 7382–7387.
- (31) Scuseria, G. E.; Schaefer, H. F., III. *J. Chem. Phys.* **1989**, *90*, 3700–3703.
- (32) Dunning, T. H., Jr. *J. Chem. Phys.* **1989**, *90*, 1007–1023.
- (33) Woon, D. E.; Dunning, T. H., Jr. *J. Chem. Phys.* **1995**, *103*, 4572–4585.
- (34) Dalton, a molecular electronic structure program, Release 2.0, 2005. <http://dirac.chem.sdu.dk/daltonprogram.org/> (accessed November 2011).
- (35) Pulfer, M.; Hu, C.-H.; Chong, D. P. *J. Chem. Phys.* **1997**, *216*, 91–98.
- (36) Helgaker, T.; Klopper, W.; Koch, H.; Noga, J. *J. Chem. Phys.* **1997**, *106*, 9639–9646.
- (37) Sæthre, L. J.; Siggel, M. R. F.; Thomas, T. D. *J. Electron Spectrosc. Relat. Phenom.* **1989**, *49*, 119–137.
- (38) Goerigk, L.; Grimme, S. *J. Chem. Theory Comput.* **2011**, *7*, 291–307.
- (39) Sæthre, L. J.; Berrah, N.; Bozek, J. D.; Børve, K. J.; Carroll, T. X.; Kukk, E.; Gard, G. L.; Winter, R.; Thomas, T. D. *J. Am. Chem. Soc.* **2001**, *123*, 10729–10737.
- (40) The measurements²⁴ were made for an equilibrium mixture of gauche and anti conformers, and the contribution from each conformer was not resolved. To facilitate a comparison to the experimental numbers, the theoretical result is therefore reported as a weighted average of values for the two conformers, using weight factors (36% gauche and 64% anti) obtained from experimental studies.⁴⁴
- (41) Holme, A.; Sæthre, L. J.; Børve, K. J.; Thomas, T. D. To be published.
- (42) Holme, A.; Børve, K. J.; Sæthre, L. J.; Thomas, T. D. To be published.
- (43) Myrseth, V.; Børve, K. J.; Wiesner, K.; Bässler, M.; Svensson, S.; Sæthre, L. *J. Phys. Chem. Chem. Phys.* **2002**, *4*, 5937–5943.
- (44) (a) Verma, A. L.; Murphy, W. F.; Bernstein, H. J. *J. Chem. Phys.* **1974**, *60*, 1540–1544. (b) Bradford, W. F.; Fitzwater, S.; Bartell, L. S. *J. Mol. Struct.* **1977**, *38*, 185–194. (c) Durig, J. R.; Compton, D. A. C. *J. Phys. Chem.* **1979**, *83*, 265–268. (d) Compton, D. A. C.; Montero, S.; Murphy, W. F. *J. Phys. Chem.* **1980**, *84*, 3587–3591. (e) Heenan, R. K.; Bartell, L. S. *J. Chem. Phys.* **1983**, *74*, 1270–1274. (f) Rasanen, M.; Bondybey, V. E. *J. Chem. Phys.* **1985**, *82*, 4718–4719. (g) Herrebout, W. A.; van der Veken, B. J.; Wang, A.; Durig, J. R. *J. Phys. Chem.* **1995**, *99*, 578–585.

A Benchmark Test Set for Alchemical Free Energy Transformations and Its Use to Quantify Error in Common Free Energy Methods

Himanshu Paliwal and Michael R. Shirts*

Department of Chemical Engineering, University of Virginia, Charlottesville, Virginia 22904-4741, United States

S Supporting Information

ABSTRACT: There is a significant need for improved tools to validate thermophysical quantities computed via molecular simulation. In this paper we present the initial version of a benchmark set of testing methods for calculating free energies of molecular transformation in solution. This set is based on molecular changes common to many molecular design problems, such as insertion and deletion of atomic sites and changing atomic partial charges. We use this benchmark set to compare the statistical efficiency, reliability, and quality of uncertainty estimates for a number of published free energy methods, including thermodynamic integration, free energy perturbation, the Bennett acceptance ratio (BAR) and its multistate equivalent MBAR. We identify MBAR as the consistently best performing method, though other methods are frequently comparable in reliability and accuracy in many cases. We demonstrate that assumptions of Gaussian distributed errors in free energies are usually valid for most methods studied. We demonstrate that bootstrap error estimation is a robust and useful technique for estimating statistical variance for all free energy methods studied. This benchmark set is provided in a number of different file formats with the hope of becoming a useful and general tool for method comparisons.

1. INTRODUCTION

Simulation and theory communities have developed substantial interest in using free energy calculations for molecular design problems. Specifically, free energy calculations can guide experimental screening techniques for measuring biological interaction energies and offer the potential of a faster and cheaper way to get thermodynamic information over large chemical spaces in a variety of molecular contexts.¹ For example, drug design² requires prediction of binding affinities, tautomers, protonation states, membrane permeabilities, and solubilities, all of which involve free energy calculations.^{3–6} Similarly, free energy calculations could become useful tools in material design problems ranging from improved protein selectivity and stability on chromatographic surfaces⁷ to tailoring metal organic frameworks⁸ for applications, such as gas storage and separation. Such potential uses extend to the design of new nanomaterials, such as therapeutic dendrimers, heteropolymers, and hyperbranched polymers for molecular recognition, imaging, sensing/signaling, and controlled payload delivery.^{9–13} However, substantial roadblocks to routine use of molecular simulations as a complement to experiment include confusion over suitability of methods for different molecular problems and the lack of rigorous, validated understanding about the reliability of free energy calculations and other observables estimated using statistical methods.

Other computational fields have successfully benchmarked and tested computational methods to improve the reliability and thus the utility of simulations. The field of computational fluid dynamics has also grappled with issues of reliability and standardization. During the late 90s, substantial research efforts in the field of computational fluid dynamics (CFD) were focused on establishing validation benchmarks to improve the reliability of CFD simulations in various design applications.^{14–16}

This research helped to bring down costs, increase data fidelity, and reduce design cycle time in the early development phases of new airplanes,¹⁷ Formula 1 cars,¹⁸ treatment and diagnosis of cardiovascular diseases,^{19,20} and off-shore oil rigs.²¹ Similar validation benchmarks were developed for simulations in the nuclear industry, improving the reliability in nuclear reactor safety, underground nuclear waste storage, and nuclear weapon safety.²² Quantum chemists maintain validation databases for comparison between experimental methods and different QM methods.²³ For molecular simulation to play a similar role in molecular engineering design,²⁴ benchmarks and validation sets must be established. In this paper we aim to provide one set of tools for improved standardization of molecular simulations through the first version of a benchmark set for free energy calculations for molecular transformations.

There are a large number of free energy methods available,^{25–31} which by early 2011 have been cited collectively over 4600 times, with 20% of those citations in the last 18 months.³² However, the simulation field lacks consensus in choosing a method most appropriate for a given molecular design situation. At least three fundamental issues contribute to this confusion.

First, there is a lack of standard test cases for rigorous comparisons between different free energy methods. Studies of new methods and method comparisons frequently use relatively simple model systems, such as a one- or two-dimensional analytically solvable potential energy function,^{28,33} solvation of Lennard-Jones sphere,^{28,34} alchemical changes between small molecules,^{31,35} or simplified solvation models.³⁶ These test cases may not capture all of the issues encountered in actual molecular changes. Alternatively, papers comparing methods may use complicated

Received: June 11, 2011

Published: October 18, 2011

biophysical systems, such as protein–ligand binding or pK_a determination that are hard to converge, and therefore make it difficult to accurately gauge true gains in efficiency.^{37–40} Both of these extremes put limits on our ability to decide whether a given method will be useful in actual molecular design scenarios.

Second, computing thermophysical properties by molecular simulation involves stochastic sampling of molecular configurations, and all comparisons must deal with the fact that repeated independent measurements have associated statistical error; unlike in quantum mechanics, comparisons must be done on a statistical basis, and it will never be possible to converge most calculations to arbitrary levels of precision in reasonable computational time.⁴¹

Finally, direct comparisons between methods can be difficult because of the differences between simulation code bases. Free energy calculation capabilities are recent additions to most large-scale molecular simulation codes, and most codes support usually only a small subset of available free energy methods.

As a step toward helping solve these problems, we propose the first version of a molecular test set comprising realistic systems undergoing challenging molecular transformations. We then use this test set to test the efficiency and reliability of different methods for estimating free energy differences of molecular transformations from simulation. Although the molecular design applications listed in the introduction seem very different, there are features common to all free energy calculations required for these applications. All involve determining the preference of a molecule to partition between two environments and can be calculated by way of a difference in the free energies of molecular transformation between these two environments. For example, we might wish to design a solute preferentially solvated by a protein when compared to solvent (pure water) or a different complex medium (another protein), as in the case of drug design. Alternately we might design a solvent which preferentially solvates a given solute in a mixture; for example, designing ionic liquids⁴² for sequestering CO_2 . These molecular transformations primarily involve either growing or deleting atoms, changing the size or dispersion interaction between atoms, or altering partial charge on mutation sites. Any benchmark test set must include examples of these transformations which are simultaneously challenging enough to push new methods and yet possible to evaluate with sufficiently high precision that we can reach meaningful comparisons about different methods in a reasonable amount of computer time.

The most important features of any property estimation method to understand are the statistical errors inherent in the method, including both statistical bias and statistical uncertainty, and the reliability of the method's estimate of the property of interest. Without knowledge of such features of the methods, we cannot sufficiently trust our calculations or compare two different calculations for validation purposes.

Studies commissioned by US science funding agencies on future directions for simulation based engineering and science have emphasized the fundamental need for improved uncertainty verification and validation.^{43,44} Almost all estimators of statistical quantities, like methods for calculating free energies and ensemble averages have some bias, a systematic deviation from the true answer that would be obtained with perfect sampling. Additionally, computing a given observable from independently selected samples gives different estimates of the observable; this

variation is the statistical uncertainty of the estimate. Most free energy methods also include estimates of this statistical uncertainty. However, these uncertainty estimates are themselves statistical quantities, with variation from sample to sample, and must be validated.

A few valuable studies have compared^{28,31,33,45–49} multiple free energy methods but not necessarily in a systematic way. We use our proposed benchmark set to directly compare the estimated uncertainties with the sample uncertainty. We also compute the change in the mean square error as a function of the number of intermediate states and the number of samples to capture both bias and uncertainty. We also test whether the distribution of free energy estimators is indeed Gaussian, a condition usually assumed when using statistical uncertainty estimates to calculate error. Finally, we also evaluate the bootstrap method⁵⁰ as a tool for estimating statistical uncertainty, as this method can be easily implemented for all of free energy algorithms described in this paper, and indeed generally for most statistical estimates of observables.

In this paper, we first explain our proposed benchmark test set for molecular transformations and the rationale behind the molecular choices. Next, we use this set to test and compare the accuracy, precision, and reliability of 10 free energy methods. We then present a summary of the comparison of the methods, with much of the data presented as Supporting Information because of its length, and finally present our recommendations for methods for performing free energy calculations.

2. TEST SET

The systems in this benchmark set are designed to represent “alchemical” changes, or changes of molecular identity, common to most molecular design applications. Alchemical transformations frequently require the deletion or introduction of atoms and large changes in the partial charges. Changes in torsional, angle, or dihedral parameters usually result in smaller changes in phase space, as do small changes in dispersion strength atomic radius, or charge. We therefore focus on atomic introduction/deletion and large changes in partial charge.

2.1. Minimal Test System: OPLS⁵¹ UA Methane in TIP3P Water (MS). The solvation of a Lennard-Jones (LJ) sphere, representing methane, is perhaps the simplest free energy test case that can be truly defined as molecularly realistic. There are no bonds, angles, or torsions terms nor are there solute/solvent charge terms. This system represents a minimal test of whether the free energy method is at all valid or applicable for molecular systems. We examine the transformation of coupling the sphere into water, which corresponds to the solvation free energy of this molecule.

2.2. Charge Mutation System: Dipole Inversion in TIP3P Water (DI). We use an OPLS-UA ethane molecule with the addition of +1/−1 charges on the two atomic centers. This setup avoids computing free energies of ions directly, as changing the total charge of a system with periodic boundary conditions is not always handled exactly in many codes, and requires numerous complicated corrections.⁵² This test measures whether a method can handle large water rearrangements around charges and the large energy differences involved in changing large partial charges. The system is a null transform; the free energy change is zero as the final state is identical to the initial state by symmetry.

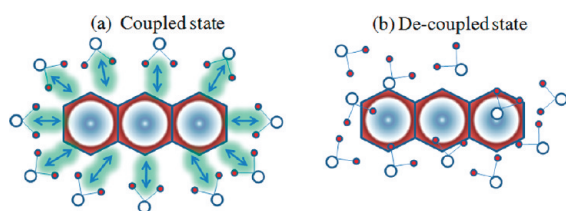


Figure 1. (a) In the coupled state or solvated state both intermolecular and intramolecular interactions for anthracene are turned on. (b) In the decoupled state or vacuum state the intermolecular interactions with water molecules are turned off.

2.3. Large Molecule Mutation System: Absolute Hydration Free Energy of UA Anthracene in TIP3P Water. In our third test, we compute the solvation of anthracene via the decoupling of the intermolecular interactions from water. This system tests whether the method can handle introduction or deletion of multiple atomic sites efficiently. Importantly, there are no internal ligand degrees of freedom to complicate the analysis. Force field parameters are taken from Pitera and Van Gunsteren.⁵³ Originally, we chose a null transformation of anthracene to anthracene via a benzene intermediate, but the simpler solvation problem was eventually chosen because of the difficulty of supporting such multiphase transformations in other codes and the difficulty of interpreting the statistics of multiple transformations; a key requirement of the benchmark set is simplicity of use. For this test set, we have used a decoupling scheme, turning off the interactions between the solute and the solvent but keeping the intramolecular interactions in the solute turned on as shown in Figure 1. The free energy change of this transformation corresponds to the desolvation free energy of anthracene, but we report the solvation free energy for ease of interpretation.

We note that although these are simplified systems, the dipole inversion and anthracene simulations are by no means toy models. The dipole moment of the dipole inversion test case is 7.20 Debye, significantly greater than the dipole moment of most small molecules, and includes a +1 to −1 charge difference on each atom. The anthracene solvation test set involves the disappearance of three aromatic rings with a total of 14 heavy atoms, which is on the high end of most molecular transformations. There are of course a number of other possible test molecules that could be examined; we will reserve discussion of future extensions of this test set for the discussion.

3. FREE ENERGY METHODS AND ERROR PROPAGATION

Using this benchmark set, we evaluated a total of 10 free energy methods, chosen specifically because all can be computed from the same set of simulation samples. In the following presentations, we assume that the simulations are performed in the isothermal–isobaric ensemble, and thus the Gibbs free energy ΔG is the quantity of interest; the Helmholtz free energy can also be computed if the simulations are performed in the canonical ensemble. U indicates the generalized potential energy, which in the case of the isobaric–isothermal ensemble is actually $U + PV$, β is $1/k_B T$, and λ is a coupling parameter connecting the initial and final states in a user-chosen manner. Brackets indicate an ensemble average over the appropriate ensemble. These methods were chosen to represent a diversity of the most commonly used methods; several of them have alternate variants, and we

have not tried to capture all possible variants. The purpose of this paper is to propose a molecular transformation test set and demonstrate its utility in method comparison, not necessarily a test of all possible free energy methods. All of the 10 methods were chosen because they can be computed from the same equilibrium samples, rather than computed independently. We have chosen not to examine expanded ensemble methods⁵⁴ or nonequilibrium methods, such as Wang–Landau⁵⁵ and Jarzynski's⁵⁶ relationship-based methods, they would require independent simulations for each method as well as each having additional parameters that would have to be chosen for optimality. However, by distributing the starting configurations and parameters in a number of formats, performing comparisons between other methods and the methods presented in this paper will hopefully become significantly easier for other research groups.

3.1. Thermodynamic Integration²⁴ Using a Trapezoid Rule (TI) and a Cubic Spline²⁶ (TI3). For TI, we compute the ensemble average of the derivative of potential energy function with respect to a coupling parameter λ for a system, i.e., $\langle\langle(\partial U(\lambda)/\partial \lambda)_{\lambda_i}\rangle\rangle$ at all λ values and the corresponding variances σ_i^2 of the $\langle\langle(\partial U(\lambda)/\partial \lambda)_{\lambda_i}\rangle\rangle$ distributions:

$$\sigma_i^2 = \langle\langle x^2 \rangle\rangle - \langle\langle x \rangle\rangle^2, \quad \text{where } x = \langle\langle(\partial U(\lambda)/\partial \lambda)_{\lambda_i}\rangle\rangle \quad (1)$$

The $\langle\langle(\partial U(\lambda)/\partial \lambda)_{\lambda_i}\rangle\rangle$ values at different intermediates are interpolated and then integrated to get an overall free energy change:

$$\Delta G_{10} = G(\lambda = 1) - G(\lambda = 0) = \int_{\lambda=0}^{\lambda=1} \langle\langle(\partial U(\lambda)/\partial \lambda)_{\lambda_i}\rangle\rangle d\lambda \quad (2)$$

For TI we have used a linear interpolation, which leads to the standard trapezoid rule to integrate the total free energy. For TI3, the $\langle\langle(\partial U(\lambda)/\partial \lambda)_{\lambda_i}\rangle\rangle$ vs λ_i curve is fit piecewise to a natural cubic spline and then integrated analytically using the coefficients of the cubic equation (see Appendix for the derivation). Both the trapezoidal and the cubic spline integration can be expressed in the form of weighted sum of individual $\langle\langle(\partial U(\lambda)/\partial \lambda)_{\lambda_i}\rangle\rangle$:

$$\Delta G_{10} = \sum_{i=1}^K W_i \langle\langle(\partial U(\lambda)/\partial \lambda)_{\lambda_i}\rangle\rangle \quad (3)$$

Here the W_i 's are the respective weights corresponding to each state and K is the total number of intermediate states. The variance σ_{10}^2 of this estimate of free energy can be calculated by the following variance propagation formula:

$$\sigma_{10}^2 = \sum_{i=1}^K W_i^2 \sigma_i^2 \quad (4)$$

Occasionally, some researchers have computed the variance of the free energy over each interval, i to $i+1$ individually, and then propagated these results into the total variance. This is incorrect, since the variance of each interval is correlated to the variance of the neighboring intervals. For example, the free energy difference between states 1 and 2 and between states 2 and 3 both contain statistical information from state 2. It is important to propagate the uncertainty directly using eq 4 to avoid potential errors.

A number of alternative TI schemes have been proposed.^{26,57,58} However, many of these schemes require some knowledge of the magnitude of the statistical uncertainty for optimality. Other schemes use nonlinear fits to two different functional forms separately

describing LJ and Coulomb contributions to the free energy.⁴⁷ Such schemes are not particularly flexible and introduce integration bias that is difficult to quantify. By using cubic splines, we can obtain a higher order formula independent of functional form of $dU/d\lambda$, while propagating error using the same formalism as is used in standard TI (eq 4). For specific applications, different TI weighting schemes may be more appropriate to the particular curvature encountered, and these application-specific methods might easily be better for their intended application than a general-purpose spline, such as the one evaluated here. Because of this large range of slight variants of the weights, we will not attempt to classify all possible TI methods in this study but include one higher order algorithm to examine some method beyond the simplest trapezoidal case.

3.2. Exponential Averaging (EXP) in Two Forms: DEXP and IEXP^{27,28}. In exponential averaging schemes, the free energy change ΔG_{ij} is calculated using the exponential average of the difference of the potential energies ΔU_{ij} between two states i and j over one of the ensembles. The free energy difference as a function of potential energy difference ΔU_{ij} and N samples is then

$$\Delta G_{ij} = -\frac{1}{\beta} \ln \left(\frac{1}{N} \sum_{n=1}^N \exp[-\beta \Delta U(x_n)_{ij}] \right) \quad (5)$$

This averaging is performed using samples from state i to compute potential energy differences ΔU_{ij} from state i to state j . The free energy of the reverse process can be computed using samples from state j and computing potential energy differences to state i . Since the labels themselves are arbitrary, to remove ambiguity in the direction we will describe such computations as being either “deletion” or “insertion”. We will call ΔU_{ij} taken in the direction of decreasing entropy as an “insertion” step and ΔU_{ij} taken in the direction of increasing entropy as a “deletion” step, as inspired by Wu et al.⁵⁹ Hence the free energy method with ΔU_{ij} stepping in the direction of increasing entropy in eq 5 is labeled as deletion exponential averaging (DEXP), and the free energy method with steps of ΔU_{ij} in the direction of decreasing entropy in eq 5 is labeled as insertion exponential averaging (IEXP). In both cases, the variance σ_{ij}^2 between two adjacent intermediate states can be estimated using standard point estimation theory as

$$\sigma_{ij}^2 = \frac{1}{N} \left(\frac{\sigma_x}{\langle x \rangle} \right)^2, \quad x = \exp[-\beta \Delta U(x_n)_{ij}] \quad (6)$$

In both the exponential averaging methods the overall free energy change ΔG_{10} is the sum of intermediate free energy changes ΔG_{ij} , and so the variance σ_{10}^2 is simply the sum of the associated variances σ_{ij}^2 . In some cases, the changes from the i state to the $i-1$ state and $i+1$ states might both be deletion or insertion cases; in this case, all the sampling performed at i and the two estimates of the free energy difference will not be statistically independent. Complicated molecular changes will frequently involve both addition and subtractions of accessible phase space and thus will fall somewhere in between these two general schemes.

Any free energy change involving inherent directionality, such as exponential averaging, requires careful definitions to ensure that the direction of entropy change remains constant throughout the process. Otherwise, we cannot interpret the

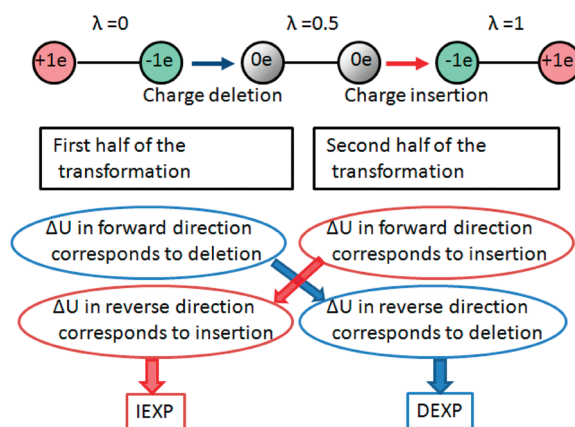


Figure 2. Free energy differences of transitions in the direction of increasing and decreasing entropy should be added separately to get the overall free energy for a dipole inversion.

entire transformation as a deletion or an insertion process. Methane and anthracene solvations involve moving molecules from vapor to liquid phase, resulting in a decrease in total entropy. However, in the dipole inversion case, we have a symmetric transformation, and thus deletion and insertion happen within a single process. In dipole inversion, going from a very large magnitude dipole to a small apolar intermediate, we have an increase of entropy as the water around the particle becomes less structured, and thus we use the term deletion. From the intermediate uncharged intermediate state to the reversed $-/+$ dipolar state during the second half of the inversion, we have the reverse process, and we use the term insertion consistent with the entropy direction definition. To use the terminology IEXP, Gaussian estimate with insertion (GINS), DEXP, and Gaussian estimate with deletion (GDEL) pathways, we therefore need to combine mixed halves of what would typically be called the forward and reverse pathways as illustrated in Figure 2. Although these particular sums are nonzero, they provide a consistent definition of the statistical variance of insertion and deletion. The statistical variance for symmetric transformations will simply be the average of the variance for the deletion and insertion processes.

3.4. Bennett Acceptance Ratio (BAR)²⁹. The Bennett acceptance ratio uses samples of the potential energy in both i to j and j to i directions to obtain a provably minimum variance estimate of the free energy difference. Calculation of the free energy change between any two intermediate states through BAR requires self-consistent solution of the two equations:

$$\Delta G_{ij} = \frac{1}{\beta} \ln \left(\frac{\sum_{k=1}^{N_j} \frac{1}{1 + \exp[-\beta(\Delta U_k^j - C)]}}{\sum_{l=1}^{N_i} \frac{1}{1 + \exp[-\beta(\Delta U_l^i - C)]}} \right) + C - \frac{1}{\beta} \ln \left(\frac{N_j}{N_i} \right) \quad (7)$$

$$C = \Delta G_{ij} + \frac{1}{\beta} \ln \left(\frac{N_j}{N_i} \right) \quad (8)$$

The first equation is true for any constant C , but when eqs 7 and 8 are solved self-consistently the ΔG_{ij} will have minimized variance. There exists a large number of ways to solve the equations self-consistently, and a complete discussion of the

best methods is beyond the scope of this paper. The variance σ_{ij}^2 in ΔG_{ij} for any C can be estimated as

$$\sigma_{ij}^2 = \frac{1}{\beta^2 N_i} \left[\frac{\langle f^2(x) \rangle_i}{\langle f(x) \rangle_i^2} - 1 \right] + \frac{1}{\beta^2 N_j} \left[\frac{\langle f^2(x) \rangle_j}{\langle f(x) \rangle_j^2} - 1 \right] \quad (9)$$

where $f(x)$ is the Fermi function $1/(1+x)$ and $x = \exp[\beta(\Delta U - C)]$. The total free energy change is the sum over changes between consecutive intermediate states. Typically, the variance in the full free energy is computed by assuming independent error and summing the variance for consecutive intermediate states. However, the assumption that the errors add independently is not correct, since the free energy difference from $i-1$ to i and from i to $i+1$ states both depend on the potential energy at i , so their variances are not independent. There is thus no general formula to obtain a statistically unbiased estimate of an entire transformation computed by a series of BAR calculations between neighboring states.

3.5. Unoptimized Bennett Acceptance Ratio (UBAR). Equation 9 in Section 3.4 is valid for any initial estimate of the free energy, though choices of C not given by the implicit equation (eq 9) will not have minimum variance. If we make the choice of $C = \beta^{-1} \ln(N_j/N_i)$, we no longer need to self-consistently solve equations. This can avoid saving, reading, and reprocessing all of the data, potentially saving significant disk space or memory, at a cost of decreased statistical efficiency and increased bias. We can instead accumulate the averages in eqs 7–9 as the simulation progresses. If each intermediate free energy is relatively near zero, then this free energy estimate will be close to optimal. This estimator is directly equivalent to the minimum variance version of transition state Monte Carlo, where Barker acceptance probability⁶⁰ is used.⁶¹

3.6. Range-Based Bennett Acceptance Ratio (RBAR). If we keep track of the ensemble averages in eqs 11 and 12 for a range of trial values of C , we will obtain a number of estimates of the best estimate free energy.⁶¹ Of these free energy estimates, the one that corresponds most closely to the input value of the free energy in the formula for C in eq 8 will be the least biased and will have minimum variance. By choosing this particular value of the free energy from the range of values, we are essentially precalculating the self-consistent solution. To apply this method, a range of starting values of C is chosen. This trial C is fed as an initial guess, and C is calculated using a single iteration of eq 11, with corresponding ΔG and σ then calculated. Accumulated averages are maintained for each choice of C . A decent estimate of the range of C (and therefore ΔG) is therefore a requirement for using this method. In some cases, it may end up being more costly than BAR, as accumulated averages must be maintained for a certain number of trial free energy values, instead of simply performing 5–10 self-consistent iterations. However, the advantage of what we will call in this paper RBAR is that data from each simulation step does not need to be retained for postprocessing, as is required with BAR, and only the accumulated averages need to be retained.

3.7. Multistate Bennett Acceptance Ratio (MBAR)³⁰. MBAR is a method to find the free energies of K states simultaneously by minimizing the $K \times K$ matrix of variances of the free energy differences of these K states simultaneously. The derivation of MBAR is a straightforward if mathematically difficult extension of the derivation BAR to more than two states considered simultaneously. This can be a significant improvement

over BAR, which minimize variances of the free energy differences for two states at a time. For MBAR, the equation:

$$G_i = -\frac{1}{\beta} \ln \sum_{k=1}^K \sum_{n=1}^{N_k} \frac{\exp[-\beta U_i(x_{kn})]}{\sum_{k'=1}^K N_{k'} \exp[\beta G_{k'} - \beta U_{k'}(x_{kn})]} \quad (11)$$

is solved self-consistently for each G_i . $\Delta G_{ij} = G(\lambda_j) - G(\lambda_i)$ gives the free energy change between two states i and j . The statistical variance of ΔG_{ij} , σ_{ij}^2 is calculated using eqs 8 and 12 in the paper by Shirts and Chodera.³⁰

Importantly, the popular weighted histogram analysis method (WHAM)⁴⁶ for computing free energies, based on the multiple histogram algorithms of Ferrenberg and Swendsen,^{62,63} can be seen as a histogram approximation to this equation. If instead of computing sums of the samples, we bin the energies U_i into a histogram for each of the intermediate states, then the MBAR equations become equivalent to WHAM equations. Similarly, if one reduces the histogram width to zero, one arrives at the MBAR equations,⁶⁴ though this derivation does not allow one to calculate an error estimate. Thus, by testing MBAR, we are also testing WHAM in the limit of sufficiently narrow histograms.

3.8. Gaussian Estimate of Exponential Averaging in Two Forms: GDEL and GINS³¹. If $\sigma_{\Delta U}^2 = \langle \Delta U^2 \rangle - \langle \Delta U \rangle^2$ is finite and we approximate the ΔU_{ij} distribution as a Gaussian, the free energy can be expressed as a sum of moments of the probability distribution of energy differences³⁶ by

$$\Delta G_{ij} = \langle \Delta U \rangle_{ij} - \frac{\beta}{2} \sigma_{\Delta U_{ij}}^2 \quad (12)$$

The variance over N samples of this free energy difference is approximated in the limit of no higher moments by

$$\sigma_{ij}^2 = \frac{\sigma_{\Delta U_{ij}}^2}{N} + \frac{\beta^2 \sigma_{\Delta U_{ij}}^4}{2(N-1)} \quad (13)$$

If the distribution ΔU_{ij} is close to Gaussian, then this estimation method can minimize the statistical effect of rare events, resulting in a more efficient and substantially simpler estimate method. To remove ambiguities with respect to direction of the process, we use the same convention of deletion and insertion as described for exponential averaging. In eq 12, when ΔU_{ij} is in the direction of increasing entropy, we refer to this as GDEL, and if we use ΔU_{ij} in the direction decreasing entropy, we refer to this estimate as GINS. Summing the free energy changes between intermediates again gives the total free energy changes. Total variance is calculated assuming independent sampling at each state, which is not an approximation here, as each calculation depends on samples from only one state. The total free energy is calculated by summing over the free energy changes between neighboring states.

4. SYSTEM PREPARATION, SIMULATION PARAMETERS AND STATISTICAL TESTS

4.1. System Preparation and Simulation Parameters. Topologies for united atom (UA) methane, dipole inversion, anthracene solvation test systems were created by a combination of automated tools (Dundee PRODRG,⁶⁵ OpenEye libraries⁶⁶) and manual editing and are available on the Alchemy.org Web site, <http://www.alchemy.org>. Starting configurations were

generated using GROMACS 4.0.4. The automated topologies were solvated using GROMACS genbox, and these solvated systems were minimized with the low-memory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS)⁶⁷ minimization method, followed by steepest descent minimization. All systems were then equilibrated at constant volume at 300 K for 100 ps, using Langevin dynamics with a time step of 0.002 fs.^{68,69} All hydrogen-containing bonds were constrained using the SHAKE algorithm to a relative tolerance of 10^{-12} . The systems were then equilibrated at constant pressure at 1 atm using a Parrinello–Rahman barostat⁴⁴ and a Nose–Hoover thermostat⁴⁵ for 100 ps. A coupling time constant of 5 ps was used for both thermostat and barostat. A switching function was used for both particle mesh Ewald (PME) and van der Waals potentials. The PME switch started at 0.88 nm with a coulomb cutoff distance of 0.9 nm for electrostatics. Other PME parameters were: Fourier spacing of 0.12 nm, fourth-order B-spline interpolation, and a Ewald tolerance of 10^{-8} . A van der Waals switch at 0.8 nm and cutoff distance of 0.9 nm were used. A long-range van der Waals dispersion correction was used for both energy and pressure.

4.2. λ Values and Spacing between Intermediate States for Free Energy Calculations. In order to examine the change in bias, statistical error, and mean square error as a function of the spacing between coupling parameter λ values, we choose two sets of λ states for each model: a full λ set, and a sparse λ set.

4.2.1. Full λ Set. Initial simulations (5 ns long, including 0.5 ns equilibration) were performed with 21 equally spaced λ values to guide the selection of the λ values for the main study. The free energy analysis was done using TI. Intermediate states were chosen so that each window contributed equally to the total error such that the uncertainty $\delta(\Delta G_{i,i+1})$ vs λ_i curve for TI was flat, specifically ensuring that the maximum variance among all windows was no larger than the twice of the variance among all windows. The λ values were chosen such that each λ window contributed 0.027 ± 0.006 kJ/mol to the total uncertainty for methane solvation, 0.047 ± 0.005 kJ/mol to the total uncertainty for dipole inversion, and 0.029 ± 0.011 kJ/mol to the total uncertainty for anthracene solvation. For UA methane solvation (MS), 8 intermediates were selected: $\lambda = [0.0, 0.2, 0.4, 0.5, 0.6, 0.7, 0.8, 1.0]$, where $\lambda = 0$ denotes fully interacting UA methane in water and $\lambda = 1$ denotes “ghost” UA methane, where there are no interactions with the solvent. For consistency of sign between test cases, the reported results are in terms of the reverse process, the solvation of methane. For dipole inversion (DI), we include 11 intermediate states: $\lambda = [0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0]$. Here $\lambda = 0$ denotes a starting $+/-$ configuration of the dipole; $\lambda = 1$ denotes the reversed configuration of dipole, i.e., $-/+$, with $\lambda = 0.5$ a state with zero partial charges. For anthracene solvation (AS), the full set contains 15 total states, with $\lambda = [0.0, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, 1.0]$; $\lambda = 0$ denotes fully interacting UA anthracene in water; $\lambda = 1$ denotes the vacuum-state anthracene with no interactions with water. Again this corresponds to a desolvation process, and in tables and charts, we report the free energy of hydration, which is simply the reverse process and so includes a sign reversal. For RBAR, this spacing means that the largest free energy between intervals is approximately 35 kJ/mol, meaning we must use range of -40 to 40 kJ/mol, and we choose increments of 1 kJ/mol.

4.2.2. Sparse λ Set. For methane solvation, we chose only three λ states $[0, 0.5, 1]$. For dipole inversion, the sparse set was generated by picking every alternate λ along with $\lambda = 0.5$ (which represents zero net charge) $[0, 0.2, 0.4, 0.5, 0.6, 0.8, 1.0]$,

reducing the number of states from 11 to 7. For the AS test set, every third λ was chosen to create the sparse λ set $[0.0, 0.2, 0.5, 0.7, 0.85, 1.0]$, reducing the number of states from 15 to 6. Note that we did not need to run the separate simulations for the sparse states; we merely select for analysis only a subset of the states from the full λ set. For RBAR, this spacing means that the largest free energy between intervals is approximately 66 kJ/mol, meaning we must use a range of -70 to 70 kJ/mol, again with increments of 1 kJ/mol. In order to also test the bias with respect to number of states, for each model we conduct a set of 5 ns simulations at 51 equally spaced states, with a spacing 0.02 between 2 neighboring λ states.

4.3. Generating an Ensemble of Uncorrelated Configurations. The most important use of the benchmark test set in this paper is to compare estimates of the statistical error of different estimators of the free energy with direct sample error obtained by repeating the experiment N times. For this purpose, we started by generating 100 uncorrelated starting configurations. Using the $\lambda = 0$ state from the 5 ns test runs, we used the GROMACS program `g_analyze` to compute the autocorrelation time of the potential energy, kinetic energy, total energy, Coulomb interactions, and derivative $dU_{\text{pot}}/d\lambda$ for all three systems. We used block averaging⁷⁰ using the GROMACS `g_analyze` program to compute the autocorrelation times. The autocorrelation time of potential energy was chosen since it was the longest correlation time of the observables listed here for all molecules. The autocorrelation times of potential energy for UA methane solvation, dipole inversion, anthracene solvation were 25, 30, and 25 ps, respectively.

To generate initial configurations, each of the three test systems were run for 20 ns using a prerelease version of GROMACS 4.5 also used for subsequent free energy configurations. We then selected configurations separated by 2 ns as our uncorrelated starting points. The 2 ns spacing is more than 50 times longer than the 30 ps autocorrelation time of the potential energy. From each of these 10 parent configurations, 20 ns simulation were run with different random seeds to generate a new Maxwell–Boltzmann velocity distribution, and configurations were again selected every 2 ns, giving a total of 100 uncorrelated starting configurations for each system. The velocity Verlet integrator was used with a Nose–Hoover thermostat and the Martyna–Tuckerman–Tobias–Klein (MTTK)⁷¹ barostat was used to control temperature and pressure, respectively, with other parameters set to the defaults discussed above.

The coordinates for simulations at the first state, at $\lambda = 0$, were selected from one of the 100 uncorrelated starting configurations. The starting coordinates for each subsequent intermediate state simulation were generated by running consecutive short 10 ps equilibration runs from the ending configuration of the previous λ state. After this initial equilibration round, 5 ns of NPT simulation were then performed for each initial configuration and each separate λ state. Data from the first 500 ps were discarded as equilibration. The remaining 4.5 ns of equilibrium data for each model at each intermediate state were used for all subsequent calculations.

4.4. Statistical Tests. **4.4.1. Quantifying Accuracy and Precision in Uncertainty Estimate of an Estimator.** We estimate the statistical uncertainty for each free energy estimator in three ways. First, we compute the sample standard deviation from ΔG 's computed from the series of 100 uncorrelated simulation runs described above. Second, we compute the analytical estimates of error corresponding to each of the

methods. Finally, we use the bootstrap estimator for the standard deviation.⁵⁰

4.4.1.1. Sample Standard Deviation. To compute the sample standard deviation, we take the simulations started from the 100 initial configurations and compute free energy differences from each simulation to obtain a distribution of free energy differences. We then directly compute the sample standard deviation corresponding to each individual estimator from

$$\sigma(\Delta G) = \sqrt{\frac{\sum_{i=1}^N (\langle \Delta G \rangle - \Delta G_i)^2}{N-1}} \quad (14)$$

where $N = 100$ and $\langle \Delta G \rangle$ is the mean over the 100 values of ΔG_i 's. Crucially, the standard deviation computed from a finite-sized sample is itself a statistical quantity and must therefore have an associated uncertainty. Rigorously, in order to compute the sample standard deviation of the uncertainty $\delta(\sigma(\Delta G))$, we would need to repeat our 100 simulation experiment 100 times. Instead, we have used the bootstrap method (described in the bootstrap estimate section) to estimate $\delta(\sigma(\Delta G))$; as will be seen later, the bootstrap method is an effective way to compute estimates from independent free energy calculations. From this exercise we finally get $\langle \Delta G \rangle$ and $\langle \sigma(\Delta G) \rangle \pm \delta(\sigma(\Delta G))$; $\langle \Delta G \rangle$ here indicates not an ensemble average but the average over 100 repetitions.

4.4.1.2. Analytical Estimate. Each free energy estimator has an associated uncertainty estimator as discussed in previous sections, namely the square root of the estimated variance of the total free energy. From 100 uncorrelated starting configurations, we will obtain not only 100 ΔG 's, but 100 error estimates from each method's analytical uncertainty estimate. We denote the average and standard deviation of these estimated uncertainties over all 100 independent runs as $\langle \delta(\Delta G) \rangle \pm \delta(\delta(\Delta G))$ and call these the analytical uncertainty estimate and the standard deviation of the analytical uncertainty estimate.

4.4.1.3. Bootstrap Estimate. For each of the 100 independent free energy calculations, we also calculate a bootstrap error estimate, a well-known and robust technique in the statistical literature.⁵⁰ The bootstrap error is constructed as follows: From each set of potential energy differences or $dU/d\lambda$ values, we generate N bootstrap sets from the original set of molecular simulation data. To generate a bootstrap set, we first subsample the data using an estimate of the autocorrelation time to obtain N statistically uncorrelated values. For each bootstrap set, we draw N samples with replacement from the original set of uncorrelated measurements. For example, if our set was the integers $\{3,6,8,9\}$, then a bootstrap set would consist in randomly selecting each of the four numbers for times; $\{3,3,8,9\}$, $\{9,6,6,3\}$, and $\{8,8,8,8\}$ would all be valid sets, though clearly the last one would be the rarest. This subsampling process is then repeated many times; in this particular study, we draw 200 bootstrap sets. Standard rules of thumb suggest using 50–200 bootstrapped sets to get robust estimates of uncertainty,⁵⁰ though anecdotally some users say they get more consistent results using 1000 or more bootstrap sets. For each of these 200 bootstrap sets, we compute the free energy and the uncertainties using the estimators as if they were the original data set. This gives us 200 ΔG 's, one for each of the 200 bootstrapped sets. The average of these 200 bootstrapped ΔG 's gives $\langle \Delta G \rangle_{\text{bs}}$. The bootstrap estimate of the error, $\delta(\Delta G)_{\text{bs}}$ is the sample standard deviation of the 200 $\langle \Delta G \rangle_{\text{bs}}$ values for each of the initial configurations. The average $\langle \delta(\Delta G)_{\text{bs}} \rangle$ over all 100

initial configurations is the bootstrap error estimate. The statistical uncertainty of this bootstrap uncertainty estimate is estimated by computing the sample standard deviation over the 100 $\delta(\Delta G)_{\text{bs}}$ and is denoted by $\delta(\delta(\Delta G)_{\text{bs}})$.

The analytical estimate of a free energy estimator by definition should agree with the sample standard deviation. If it does not, then any estimate of error using the analytical estimate will be unreliable. For example, the analytic estimate of the uncertainty of EXP diverges from the true estimate well before the error in EXP itself.²⁸ If the statistics are well-behaved, the bootstrap estimate of the statistical uncertainty should also agree with the sample estimate of the uncertainty. The smaller the difference between the direct sample standard deviation error estimate $\langle \sigma(\Delta G) \rangle$ and the analytical error $\langle \delta(\Delta G) \rangle$ or bootstrap estimates, the better we know the method's variability without having to run multiple trials. Additionally, if we can show that the bootstrap estimate agrees with the sample estimate of the uncertainty, then bootstrap error can substitute for sample uncertainty estimates even when the analytical estimate fails.

4.5. Quantifying Bias of Free Energy Estimates. The average estimate from a statistically biased estimator, even if repeated many times, will still deviate from the true estimate by the bias. There are typically two types of bias in the free energy estimators considered here. Asymptotically unbiased estimators have bias with finite number of samples but the bias decreases to zero in the limit of large numbers of samples. An example is the naïve estimator of the variance in the average of a set of numbers, which is $\text{var}(a) = N^{-1} \Sigma[(\langle A \rangle - A_i)^2]$. This estimate can be shown to always be slightly too large, by an amount proportional to the number of samples, and is thus asymptotically biased. If N is replaced by $N - 1$, however, the estimator becomes unbiased for any number of samples. In the limit of very large N , the bias of an asymptotically biased estimator will be effectively zero, but a given estimator might require a very large number of samples to reach this point. Thermodynamic integration does not have asymptotic bias, because at each intermediate state, the simple average of $dU/d\lambda$ is unbiased for any numbers of samples. Exponential averaging, BAR, and MBAR are only asymptotically unbiased, though the bias of exponential averaging is usually significantly higher than that of BAR and MBAR²⁸ and careful design of the pathway can minimize this large bias to some extent.⁷² However, unlike the simple case of the estimator for the variance of the simple mean of samples, there exist no known unbiased versions of these free energy estimators.

Bias also occurs due to using a limited number of intermediate states because of lack of either phase space overlap between intermediates, as occurs in acceptance ratio methods and exponential averaging, or because of numerical integration error in TI methods. Even for a large benchmark set like the present study, computational expense and storage limits make it very difficult to approach the number of sample limit and the number of intermediate states limit simultaneously. Therefore, we attempt to estimate the two contributions to bias independently. For asymptotic bias, we compare the results from combining a fixed amount of data in either one large data set with averaging the results over or a series of shorter sets of the same data. For bias as a function of number of intermediate states, we vary the number of intermediates with fixed total length of simulation to investigate bias as a function of number of intermediate states.

4.5.1. Bias Due to Number of Samples. Data from all 100 5 ns runs is "stitched" into a single large data set analyzed simultaneously.

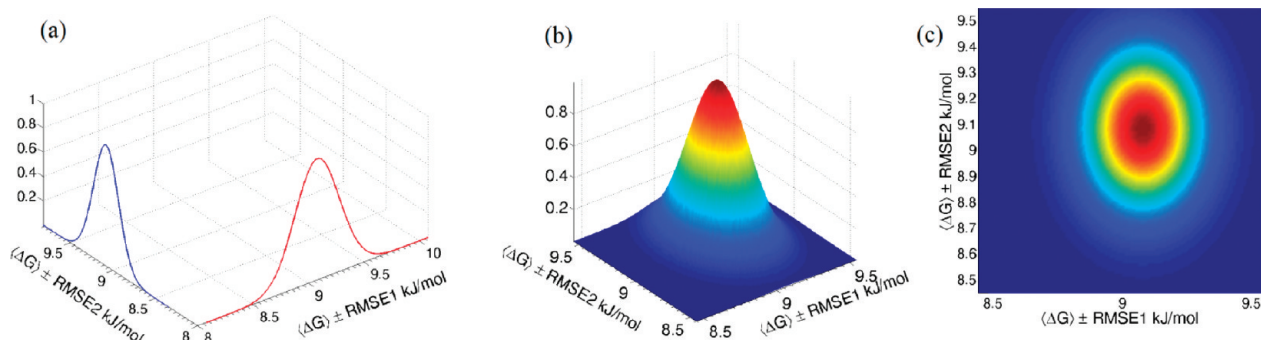


Figure 3. (a) Gaussians are plotted on mutually perpendicular axes. Both have the $\langle \Delta G \rangle$ calculated using a method (here TI for methane solvation) as mean and RMSE1 and RMSE2 as their standard deviations. (b) These are fused to generate a bivariate Gaussian plot. (c) Top view of the bivariate Gaussian plot.

The data corresponding to first 0.5 ns equilibrations were not included while estimating free energies, resulting in 450 ns total simulation data. The difference between free energy estimates computed with 450 ns of data $(\Delta G)_{450}$, and the same data used to compute an average over 100 4.5 ns trajectories $\langle \Delta G \rangle$ is what we report the bias due to number of samples.

4.5.2. Bias Due to Number of Intermediate States. We also ran a set of simulations with 51 λ states for each model, as discussed above. We can estimate the bias due to the number of intermediate states by comparing the differences between free energy estimates computed using 51 λ states $(\Delta G)_{51}$, ΔG estimated using the full set of λ states, and ΔG estimated using the sparse λ states. Besides having low and consistent error estimates, ideal free energy estimation methods should show little or no bias in these tests. In many cases, we are limited by the statistical uncertainty in determining the bias with high accuracy, since it is computationally too demanding to generate 500 ns of simulation for all 51 states for all test sets. In these cases, we can only determine if the bias is statistically insignificant with respect to the statistical uncertainty. Asymptotic bias scales with $1/N$ while statistical uncertainty scales with $N^{-1/2}$, so statistical variance is usually dominant and will always be dominant if sufficient samples are collected.

4.6. Quantifying Reliability of a Free Energy Estimator.

Root mean square error (RMSE) is the square root of the sum of squared differences between the sample and the true answer. Alternatively, we can write it as the sum of the variance estimate (σ^2) and the square of the bias. Mean square error is therefore the most general overall measure of the reliability of any statistically estimated observable. Calculating a true RMSE, of course, requires collecting an infinite number of samples at an infinite number of intermediate states to obtain the true answer. Since it is computationally impossible to reach this limit and computationally expensive to approach it, we use the free energy estimate ΔG_{450} from the 450 ns run to approximate the unbiased limit of free energies given a fixed set of intermediates and the free energy estimate ΔG_{51} from 51 λ state simulations as the unbiased limit of the free energy estimate with respect to large numbers of intermediate states, given a fixed amount of sampling. From the two different biases generated from these two reference states, we obtain two different estimates of mean square error. Neither of these is a true RMSE because we lack the true reference answer. However, these estimates of the RMSE's capture the combined effect of the statistical error and the two different sources of bias.

We define the two estimates of mean square errors as MSE1 and MSE2:

$$\begin{aligned} \text{MSE1}_i &= \sigma_i^2 + \text{bias1}_i^2, \quad \text{where } \text{bias1}_i \\ &= (\Delta G)_{450} - (\Delta G_{\text{est}})_i \quad \text{and } 1 \leq i \leq 100 \end{aligned} \quad (19)$$

$$\begin{aligned} \text{MSE2}_i &= \sigma_i^2 + \text{bias2}_i^2, \quad \text{where } \text{bias2}_i \\ &= (\Delta G)_{51} - (\Delta G_{\text{est}})_i \quad \text{and } 1 \leq i \leq 100 \end{aligned} \quad (20)$$

RMSE1 and RMSE2 are defined as the averages over the square root of MSE1_{*i*} and MSE2_{*i*}, respectively. The errors in RMSE1 and RMSE2, $\delta(\text{RMSE1})$ and $\delta(\text{RMSE2})$, respectively, are the standard deviations over the 100 RMSE1_{*i*} and 100 RMSE2_{*i*}. With the two quantities RMSE1 and RMSE2, we can examine qualitative information about the reliability of the methods using all the information from our experiments. We plot a bivariate Gaussian with variances equal to RMSE1 and RMSE2 on mutually perpendicular axes, as shown in Figure 3a, with the analytical average of the free energy estimate $\langle \Delta G \rangle$ estimated by the method as the mean. An example bivariate Gaussian RMSE plots is shown in Figure 3b. Figure 3c shows the top view of this Gaussian. The overall spread of the rings in Figure 3c is a measure of overall reliability. A poor estimator has large and/or unequal spreads in horizontal and vertical directions, yielding a large circle or an ellipse. A good estimator has small and equal spread in both the horizontal and vertical direction. Vertical spread indicates that bias due to number of intermediate states dominates the uncertainty estimate, while horizontal spread indicates bias due to number of samples dominates the uncertainty estimate.

4.7. Validating the Gaussian Distributions of the Free Energy Differences. When we express uncertainty in the free energy estimate in terms of a single number, the variance or the statistical uncertainty, rather than as a distribution, we are implicitly assuming that the errors are well described by a normal distribution, whose spread is equal to the statistical variance. Most analytical variance methods use propagation estimates that are only rigorously true in the Gaussian limit. As long as the variances are bounded (not infinite), the central limit theorem ensures that with enough samples, variances will indeed converge to the Gaussian limit. However, for finite number of samples, this assumption must be tested, not simply assumed, or else we run the risk of underestimating the chance of large deviations (“black swans”) from the average value.

To test whether the shape of the free energy distribution is Gaussian or not, we plot histograms of the distribution of free

Table 1. Statistical Uncertainty Calculated Using Three Different Approaches (Analytical $\langle\delta(\Delta G)\rangle$, Sample Standard Deviation $\sigma(\Delta G)$, and Bootstrap $\langle\delta(\Delta G)_{bs}\rangle$) for UA Methane Solvation Using the Full λ Set^a

method	$\langle\Delta G\rangle$	$\langle\delta(\Delta G)\rangle \pm \delta(\delta(\Delta G))$	$\sigma(\Delta G) \pm \delta(\sigma(\Delta G))$	$\langle\delta(\Delta G)_{bs}\rangle \pm \delta(\delta(\Delta G)_{bs})$	% dev \pm δ (% dev)
TI	9.081	0.107 \pm 0.002	0.115 \pm 0.009	0.106 \pm 0.006	-7.1 \pm 7.9
TI3	9.008	0.111 \pm 0.003	0.119 \pm 0.012	0.110 \pm 0.007	-6.9 \pm 9.5
DEXP	8.984	0.320 \pm 0.255	0.556 \pm 0.122	0.319 \pm 0.267	-42.5 \pm 47.5
IEXP	8.928	0.104 \pm 0.002	0.110 \pm 0.011	0.103 \pm 0.006	-5.8 \pm 9.6
UBAR	8.936	0.079 \pm 0.002	0.106 \pm 0.010	0.098 \pm 0.005	-25.3 \pm 7.5
BAR	8.933	0.075 \pm 0.001	0.109 \pm 0.010	0.099 \pm 0.006	-31.3 \pm 6.3
RBAR	8.937	0.075 \pm 0.001	0.109 \pm 0.010	0.099 \pm 0.006	-31.0 \pm 6.7
MBAR	8.929	0.095 \pm 0.002	0.106 \pm 0.010	0.094 \pm 0.005	-9.9 \pm 8.6
GDEL	7.042	0.093 \pm 0.002	0.136 \pm 0.011	0.121 \pm 0.008	-31.7 \pm 5.8
GINS	1.097	0.253 \pm 0.008	0.399 \pm 0.032	0.400 \pm 0.169	-36.5 \pm 5.5

^aAll quantities are in kJ/mol. $\langle\Delta G\rangle$ is not the ensemble average but the average over 100 repetitions.

Table 2. Statistical Uncertainty Calculated Using Three Different Approaches (Analytical $\langle\delta(\Delta G)\rangle$, Sample Standard Deviation $\sigma(\Delta G)$, and Bootstrap $\langle\delta(\Delta G)_{bs}\rangle$) for UA Methane Solvation Using Sparse λ Set^a

method	$\langle\Delta G\rangle$	$\langle\delta(\Delta G)\rangle \pm \delta(\delta(\Delta G))$	$\sigma(\Delta G) \pm \delta(\sigma(\Delta G))$	$\langle\delta(\Delta G)_{bs}\rangle \pm \delta(\delta(\Delta G)_{bs})$	% dev \pm δ (% dev)
TI	2.545	0.175 \pm 0.007	0.177 \pm 0.014	0.175 \pm 0.011	-1.5 \pm 8.7
TI3	3.792	0.214 \pm 0.009	0.217 \pm 0.017	0.214 \pm 0.014	-1.5 \pm 8.7
DEXP	5.631	1.343 \pm 0.524	3.179 \pm 0.679	1.628 \pm 1.186	-57.8 \pm 18.8
IEXP	9.091	0.666 \pm 0.101	0.638 \pm 0.042	0.683 \pm 0.108	4.4 \pm 17.3
UBAR	8.954	0.344 \pm 0.018	0.358 \pm 0.024	0.351 \pm 0.024	-3.9 \pm 8.2
BAR	8.926	0.225 \pm 0.006	0.263 \pm 0.014	0.232 \pm 0.014	-14.4 \pm 4.9
RBAR	8.927	0.226 \pm 0.005	0.260 \pm 0.016	0.233 \pm 0.014	-13.2 \pm 5.6
MBAR	8.928	0.232 \pm 0.006	0.262 \pm 0.015	0.232 \pm 0.014	-11.4 \pm 5.4
GDEL	-3.833	0.112 \pm 0.004	0.189 \pm 0.014	0.200 \pm 0.016	-40.6 \pm 4.9
GINS	-1.68×10^{32}	$3 \times 10^{30} \pm 34 \times 10^{30}$	$13 \times 10^{32} \pm 10 \times 10^{32}$	$1 \times 10^{32} \pm 15 \times 10^{32}$	-99.7 \pm 2.7

^aAll quantities are in kJ/mol.

Table 3. Free Energy Estimates and Corresponding Uncertainty Estimates in the Large Number of Samples (450 ns) and Large Number of Intermediate States (51 λ states) for UA Methane Solvation^a

method	$((\Delta G) \pm \delta(\Delta G))_{450,full}$	$((\Delta G) \pm \delta(\Delta G))_{450,sp}$	$((\Delta G) \pm \delta(\Delta G))_{51,full}$
TI	9.085 \pm 0.010	2.541 \pm 0.017	8.920 \pm 0.041
TI3	9.016 \pm 0.010	3.786 \pm 0.021	8.923 \pm 0.041
DEXP	9.083 \pm 0.083	12.657 \pm 4.018	8.921 \pm 0.043
IEXP	8.932 \pm 0.010	8.986 \pm 0.074	8.928 \pm 0.040
UBAR	8.939 \pm 0.010	8.930 \pm 0.035	8.922 \pm 0.040
BAR	8.939 \pm 0.010	8.920 \pm 0.023	8.921 \pm 0.040
RBAR	8.940 \pm 0.010	8.923 \pm 0.024	8.922 \pm 0.040
MBAR	8.936 \pm 0.009	8.921 \pm 0.023	8.924 \pm 0.036
GDEL	7.048 \pm 0.012	-3.837 \pm 0.020	8.847 \pm 0.043
GINS	1.101 \pm 0.041	$-1.5 \times 10^{32} \pm 3.2 \times 10^{29}$	8.841 \pm 0.040

^aBootstrap estimates are reported as they are better than analytical estimates. All quantities are in kJ/mol. The subscript (450, full) denotes the free energy estimate for 450 ns and full λ set and (450, sp) denotes the same for sparse λ set.

energies from each method against Gaussian distributions. For each Gaussian, we use the average free energy estimate as the mean and the analytical uncertainty estimate as the standard deviation of the Gaussian for each method. To validate our analysis from the visual comparison of the curves we have calculated the p -value for the Shapiro–Wilk⁷³ test, which are used to accept or reject the null hypothesis that the ensemble of 100 free energies are drawn from a normal distribution.

5. RESULTS AND DISCUSSION

Results are presented in Tables 1–5 as well as in Figures 4–15. Tables 1 and 2 contain results of the uncertainty analysis for full and sparse λ sets only for methane solvation. Tables 3–5 contain the free energy estimates corresponding to 450 ns and 51 λ states runs, bias analysis, reliability estimates of free energy, and uncertainty predictions for UA methane solvation for full and sparse λ sets. Figures 4–7 provide comparison of the accuracy

Table 4. Bias Estimates Due to Number of Samples and Number of λ States for Full and Sparse λ Sets for UA Methane Solvation^a

method	(bias1 \pm δ (bias1)) _{450,full}	(bias2 \pm δ (bias2)) _{51,full}	(bias1 \pm δ (bias1)) _{450,sp}	(bias2 \pm δ (bias2)) _{51,sp}
TI	-0.005 \pm 0.015	0.160 \pm 0.042	0.005 \pm 0.024	-6.374 \pm 0.045
TI3	-0.005 \pm 0.015	0.088 \pm 0.042	0.006 \pm 0.030	-5.131 \pm 0.046
DEXP	-0.100 \pm 0.092	0.062 \pm 0.059	-7.044 \pm 4.021	-3.308 \pm 0.150
IEXP	-0.002 \pm 0.014	0.002 \pm 0.041	0.097 \pm 0.100	0.155 \pm 0.078
UBAR	-0.003 \pm 0.013	0.014 \pm 0.041	0.024 \pm 0.049	0.032 \pm 0.053
BAR	-0.003 \pm 0.012	0.015 \pm 0.041	0.009 \pm 0.032	0.008 \pm 0.046
RBAR	-0.005 \pm 0.012	0.013 \pm 0.041	0.004 \pm 0.033	0.005 \pm 0.046
MBAR	-0.007 \pm 0.013	0.005 \pm 0.037	0.007 \pm 0.033	0.004 \pm 0.043
GDEL	-0.003 \pm 0.015	-1.802 \pm 0.044	0.001 \pm 0.023	-12.683 \pm 0.044
GINs	-0.001 \pm 0.048	-7.741 \pm 0.047	-1.4 \times 10 ³¹ \pm 3.4 \times E30	-1.7 \times 10 ³² \pm 3.4 \times 10 ³⁰

^a All quantities are in kJ/mol. The subscript (450, full) denotes the bias estimate for 450 ns and full λ set and (450, sp) denotes the same for sparse λ set. The subscript (51, full) denotes the bias estimate for 51 λ set and full λ set and (51, sp) denotes the same for sparse λ set. Bias1 refers to bias due to number of samples. Bias2 refers to bias due to intermediate states.

Table 5. RMSEs and Statistical Uncertainties in RMSEs for UA Methane Solvation^a

method	((RMSE1) \pm δ (RMSE1)) _{full}	((RMSE2) \pm δ (RMSE2)) _{full}	((RMSE1) \pm δ (RMSE1)) _{sp}	((RMSE2) \pm δ (RMSE2)) _{sp}
TI	0.148 \pm 0.054	0.208 \pm 0.084	0.237 \pm 0.076	6.377 \pm 0.178
TI3	0.153 \pm 0.059	0.172 \pm 0.069	0.290 \pm 0.093	5.135 \pm 0.218
DEXP	0.531 \pm 0.474	0.490 \pm 0.512	7.596 \pm 2.138	4.478 \pm 1.837
IEXP	0.142 \pm 0.051	0.142 \pm 0.052	0.897 \pm 0.271	0.904 \pm 0.273
UBAR	0.121 \pm 0.054	0.121 \pm 0.055	0.473 \pm 0.147	0.473 \pm 0.148
BAR	0.120 \pm 0.055	0.120 \pm 0.056	0.330 \pm 0.103	0.330 \pm 0.103
RBAR	0.120 \pm 0.055	0.121 \pm 0.056	0.331 \pm 0.102	0.331 \pm 0.102
MBAR	0.132 \pm 0.052	0.132 \pm 0.052	0.334 \pm 0.102	0.334 \pm 0.102
GDEL	0.152 \pm 0.065	1.807 \pm 0.137	0.201 \pm 0.090	12.682 \pm 0.190
GINs	0.427 \pm 0.205	7.748 \pm 0.402	3.1 \times 10 ³² \pm 1.5 \times 10 ³³	1.5 \times 10 ³² \pm 1.5 \times 10 ³³

^a All quantities are in kJ/mol. The subscripts (full) and (sp) denotes the free energy estimates for the full λ set and sparse λ set, respectively. RMSE1 uses bias due to number of samples and sample standard deviation uncertainty estimate. RMSE2 uses bias due to number of intermediate states and sample standard deviation uncertainty estimate. Free Energy estimates are largest for GINS and GDEL in sparse λ set, while the acceptance ratio methods consistently show low RMSEs and their corresponding uncertainty in the RMSEs.

and precision in free energy and the uncertainty predictions for all 10 methods for the three test cases. Bivariate Gaussian plots presenting the analysis of reliabilities of free energy methods are presented in Figures 8–10. Figures 11 and 12 compare the actual distribution of the free energies (computed using 100 5 ns simulations with full and sparse λ sets) with Gaussians using the corresponding variance and the mean estimates of the free energy methods, along with the Gaussians of 450 ns trajectory solution and the 51 λ state simulation set for comparison. In some plots, we omit GDEL and GINS, as their errors are significantly larger than the scale of errors of the other methods in the plots. Figures 13 and 14 contain the Shapiro–Wilk test comparison for all methods for methane solvation for full and sparse λ sets. Figure 15 contains a free energy convergence comparison for all methods. Tables and figures for dipole inversion and anthracene hydration free energy test cases are presented in the Supporting Information.

5.1. Validation of Uncertainty Estimates. The first column in Table 1 is the free energy change calculated as an average over the 100 repetitions from uncorrelated configurations. The next three columns are the average of the analytical estimate of uncertainty over all repetitions $\langle \delta(\Delta G) \rangle$, the sample standard deviation $\sigma(\Delta G)$, and the average bootstrap estimate of the uncertainty

over all repetitions $\langle \delta(\Delta G)_{bs} \rangle$. The last column gives the percent deviation of the analytical estimate of uncertainty from the sample standard deviation along with the propagated error. Standard deviations $\{\delta(\delta(\Delta G)), \delta(\sigma(\Delta G)), \delta(\delta(\Delta G)_{bs})\}$ for the error estimate distributions are also reported. Importantly, the error estimates are generally predicted very consistently with the exception of the error estimates for DEXP, with a relative error of usually between 5% and 10%. Thus, the error estimates obtained by almost all methods can generally be relied on to be consistent between different repetitions. This does not necessarily guarantee that they will accurately predict the variation in the free energies obtained with different data sets but at least means the error estimates will be consistent between data sets.

We find that the analytic error estimate of most methods underestimates the true sample standard deviation, most often only slightly, but occasionally more significantly. For some of the methods, this deviation is always within either one or two standard deviations and thus likely within the statistical noise. Examining the data in Tables 1 and S2 and S3 in the Supporting Information for TI and TI3, analytical and sample estimates of uncertainty differ by 1 or less standard deviation for all but the anthracene solvation with the full λ set, where it differs by approximately 2 standard deviations and is likely therefore noise.

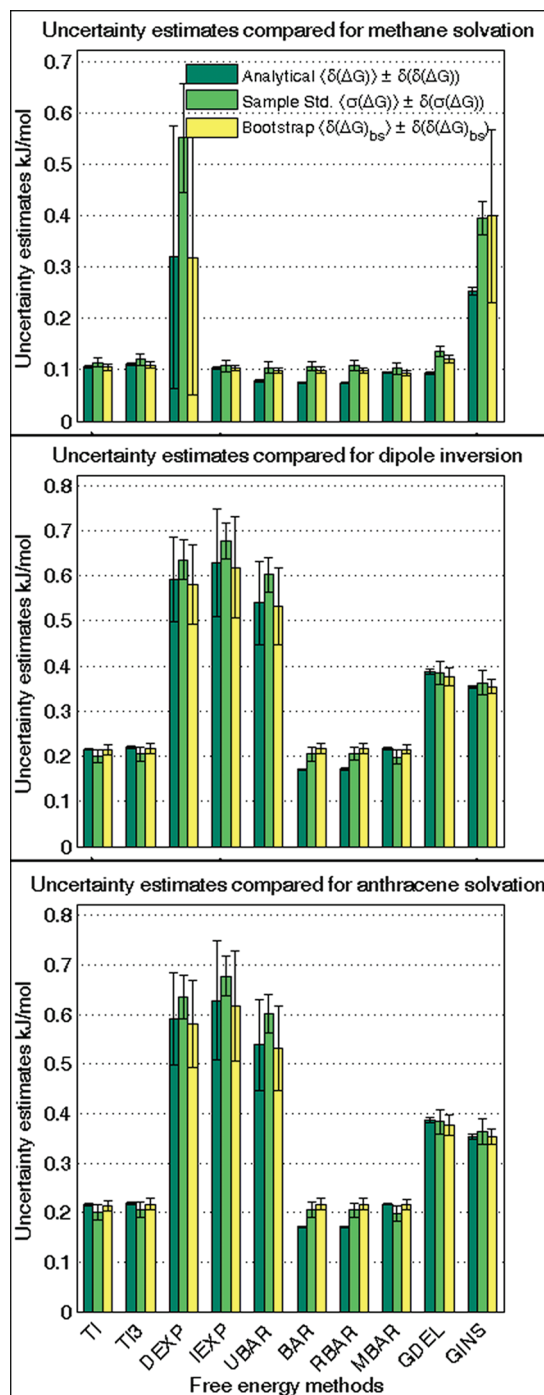


Figure 4. Uncertainty estimates (sample standard deviation, analytical, bootstrap) are plotted for all the methods in three different test cases for full λ set. Consistent free energy methods have bars equal in height, and the most precise methods have the shortest bars.

For IEXP and DEXP, analytical and sample estimates also differ by 1 or in some cases 2 standard deviations. However, analytical and sample uncertainty estimates for BAR, RBAR, and UBAR have differences of up to 5 standard deviations, which indicates that the BAR uncertainty estimates can be significantly inaccurate for multiple intermediate states. In contrast, analytical uncertainty estimates with MBAR have less than 1 standard deviation from sample uncertainty estimates. Quantitatively, as seen in

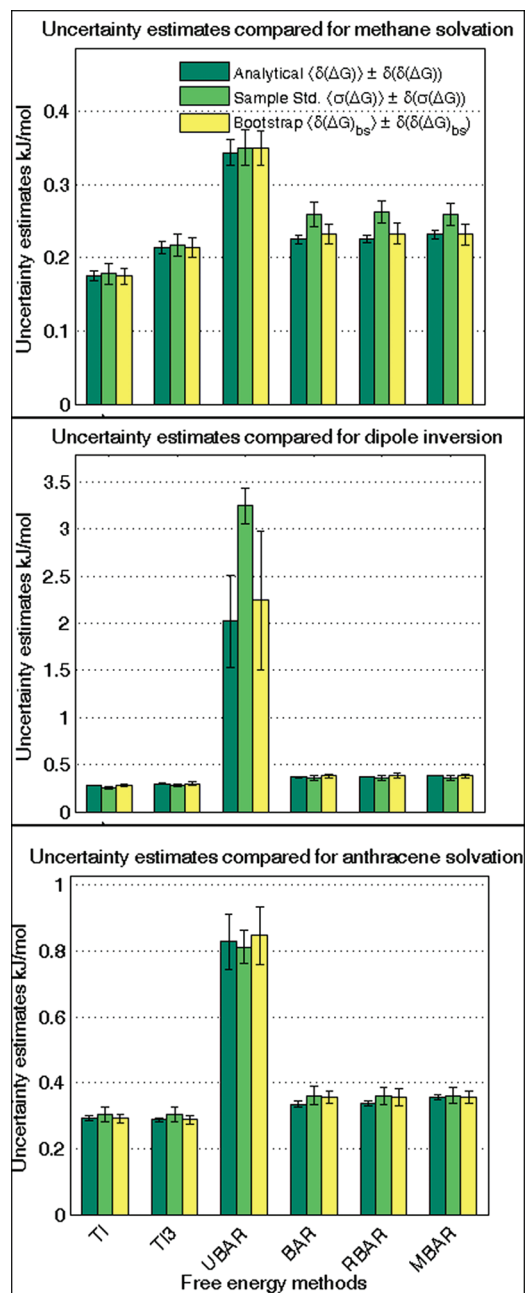


Figure 5. Uncertainty estimates (sample standard deviation, analytical, bootstrap) are plotted for six methods in three different test cases for sparse λ set. Four (DEXP, IEXP, GINS, and GDEL) are not shown because they did not converge properly for any of the three uncertainty estimates.

column six of Table 1, the percentage deviation of the analytical uncertainty estimate from sample uncertainty for TI is $-7 \pm 8\%$ and for MBAR is $-10 \pm 8\%$, with the negative sign indicating a negative deviation of the analytical estimate. In both cases, this appears to be within the noise. For BAR, the deviation from the true uncertainty is $-31 \pm 5\%$, which is clearly statistically significant, with similar results for RBAR and UBAR. As with BAR and its variants, analytical and sample estimates GDEL and GINS are off by $30 \pm 5\%$. Over all methods, MBAR and TI analytical error estimates deviate least from the sample estimate.

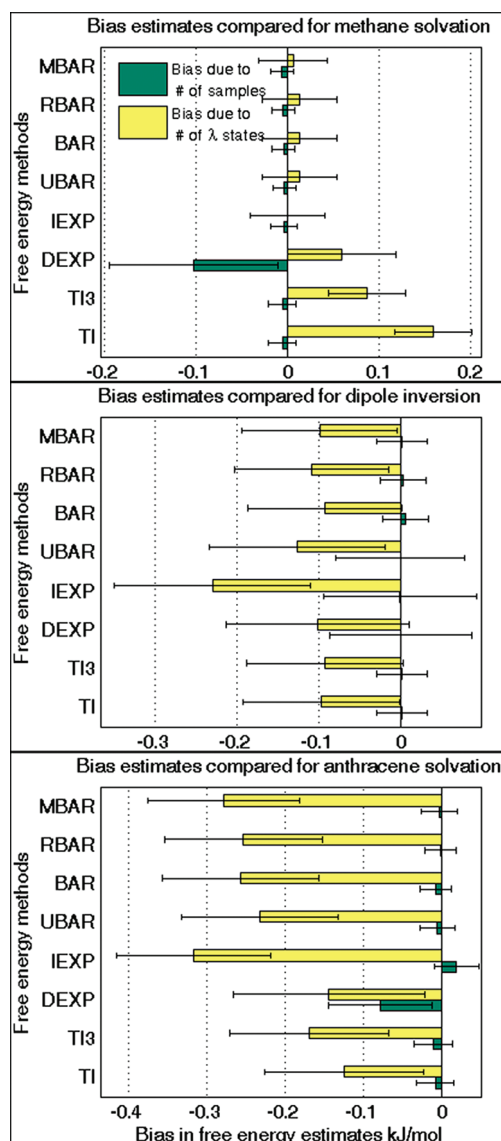


Figure 6. Bias plots for different test cases for the full λ set. DEXP, TI, and TI3 show numerically significant bias for methane, while all methods show moderate bias with respect to number of states for anthracene, with TI and TI3 showing possibly less bias.

Bootstrap uncertainty is a robust alternative to the sample standard deviation for all methods. Bootstrap estimates of the uncertainty (in column five of Table 1) are very close to sample uncertainty (in column four of Table 1). Specifically in the case of BAR, RBAR, UBAR, GDEL, and GINS bootstrap estimates of uncertainty are statistically equivalent to the sample standard error estimates, unlike their analytical counterparts. In cases where the analytical estimate does not accurately predict the sample standard deviation, such as with BAR, RBAR, and UBAR, the bootstrap method clearly provides a useful estimate of the error in the free energy without a need for performing repeated sampling.

Figure 4 visually demonstrates the efficiency of free energy methods and the consistency of uncertainty estimators. Short bars indicate precise free energy estimates. Equal height bars indicate that the analytical and bootstrap uncertainties are consistent with the sample standard deviation. For the full λ set,

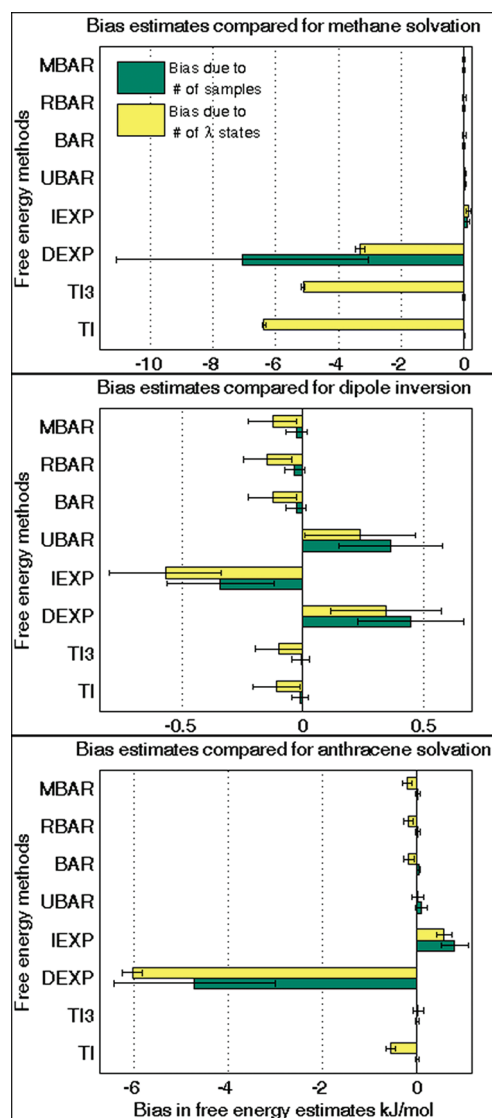


Figure 7. Bias plots for different test cases or sparse λ set. DEXP and IEXP show large biases both due to number of samples and due to number of intermediate states.

MBAR, TI, and TI3 predict free energies with the highest precision and with the most reliable error estimate. BAR, RBAR, and UBAR analytical estimates have large deviations from the standard deviation estimate, but the bootstrap estimate closely matches sample standard uncertainty estimate. IEXP and DEXP have the largest deviations of the analytical error estimate particularly for large transformations like dipole inversion and anthracene solvation. GINS and GDEL have the high uncertainty estimates and therefore give imprecise estimates of free energy.

For the sparse λ set, as shown in Tables 2 and S2 and S4 in the Supporting Information and Figure 5, TI and TI3 still show the lowest percent deviation from sample standard deviation. However, the free energy estimate of methane solvation is off by 6.5 kJ/mol for TI and 5 kJ/mol for TI3. GDEL and GINS have clearly unconverged free energy and uncertainty estimates; the free energy estimate of methane solvation for GDEL is off by 12 kJ/mol and for GINS is off by $\sim 10^{32}$ kJ/mol! This clearly indicates the failure of the Gaussian approximation of ΔU for most molecular transformation problems. The free energy

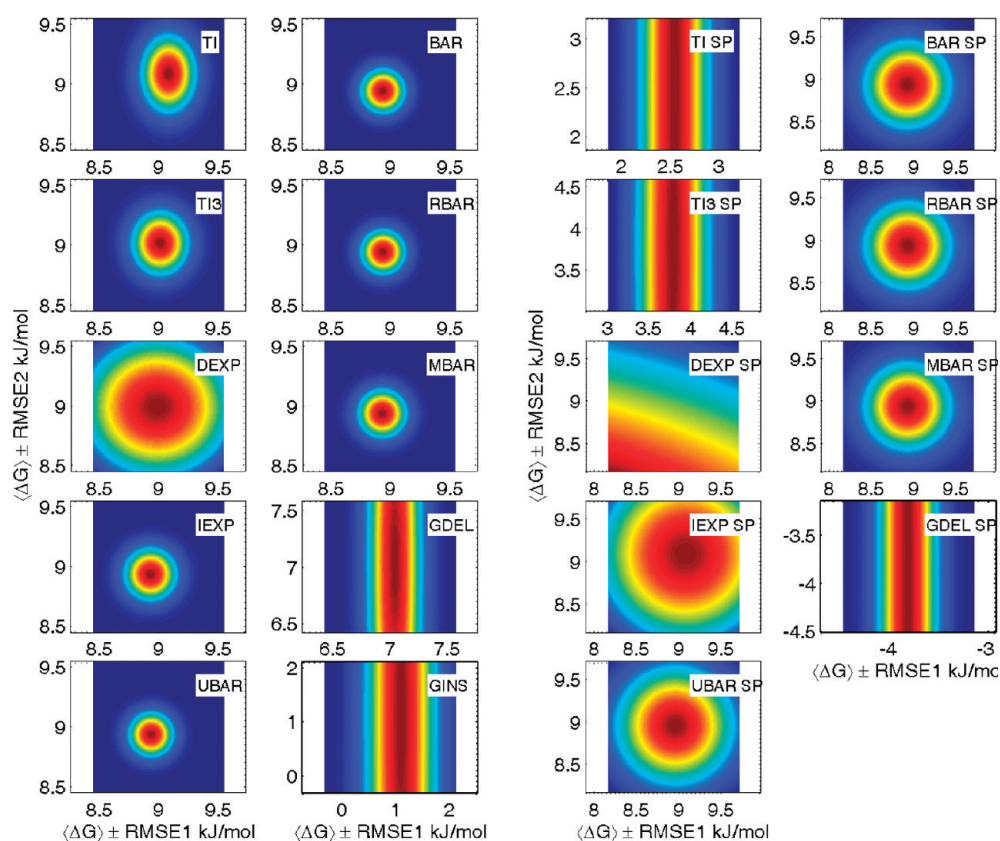


Figure 8. Bivariate Gaussian plots for UA methane solvation. TI and TI3 tend to have high bias for sparse λ sets. GDEL and GINS are not at all reliable and are not shown, and MBAR and BAR are accurate but the precision in BAR is misleading as it underestimates uncertainty, as shown in Figures 4 and 5. SP after the method name indicates sparse λ set.

estimate of methane solvation for DEXP differs from the converged answer by 3 kJ/mol, and its uncertainty estimate is 5 times larger than the largest estimated uncertainty shown in Figure 5. IEXP differs from the converged answer by only 0.2 kJ/mol, but its uncertainty estimate is twice the largest plotted uncertainty. MBAR again has the lowest and most consistent uncertainty estimates. However, unlike with the full λ set, BAR and RBAR have uncertainty estimates which are as accurate as MBAR to within statistical noise. This appears to be because as the energy differences between the i and $i + 1$ and i and $i - 1$ get larger, their correlations between the estimators computed from these energies decrease.

Bootstrap and analytical estimates of the error in the sparse λ set are slightly lower than the sample standard deviation for acceptance ratio methods for methane solvation, though not for the other two molecules. For this sparse set of λ states, MBAR does not provide the same clear advantage over BAR in estimating the uncertainty as with the full λ set. We hypothesize that this advantage may only exist when the overlap between states is non-negligible. However, even our full λ set uses relatively aggressive spacing compared to typical free energy calculations. Therefore, unless we are certain we are in a low overlap regime, MBAR will be preferred to BAR.

5.2. Analysis of Bias. The statistical uncertainty is the most important measure to quantify in order to understand the reliability of the free energy estimate but understanding systematic bias including both bias due to finite sample size and numerical integration is also important. Table 3 shows the free energy and the uncertainty estimates predicted by different methods for UA methane solvation

for large number of samples (450 ns trajectory) and large number of intermediate states (51 λ states). Tables 4 and 5 include estimates of both types of bias in free energy estimates, with respect to number of samples and with respect to number of intermediate states, and the corresponding RMSE estimates for UA methane solvation. For UA methane solvation, MBAR, BAR, RBAR, UBAR, TI, and TI3 have very low bias in free energy estimates with respect to number of samples. The acceptance ratio methods, MBAR, BAR, RBAR, and UBAR, have biases with respect to number of states within the statistical noise. TI and TI3 show larger bias in free energy estimates with respect to number of states than the other methods. DEXP and IEXP show large biases in free energy estimates both with respect to number of samples and states. GDEL and GINS show the largest bias with respect to number of states. All methods show a larger bias with respect to the number of λ states compared to the bias with respect to number of samples. However, this may be an artifact of the lower precision of bias determination as a function of the number of states.

Figures 6 and 7 show the biases for different methods for all three test cases. When error bars are larger than the bias bars, for example, for MBAR in methane solvation, comparisons for accuracy between free energy methods become difficult. DEXP and IEXP have statistically significant bias for all the cases except in methane solvation. For UA methane solvation sparse set, TI unsurprisingly has the largest bias due to number of states, as only three states were used.

Overall, MBAR, BAR, UBAR, and RBAR have consistently less bias compared to other methods and hence are more accurate for

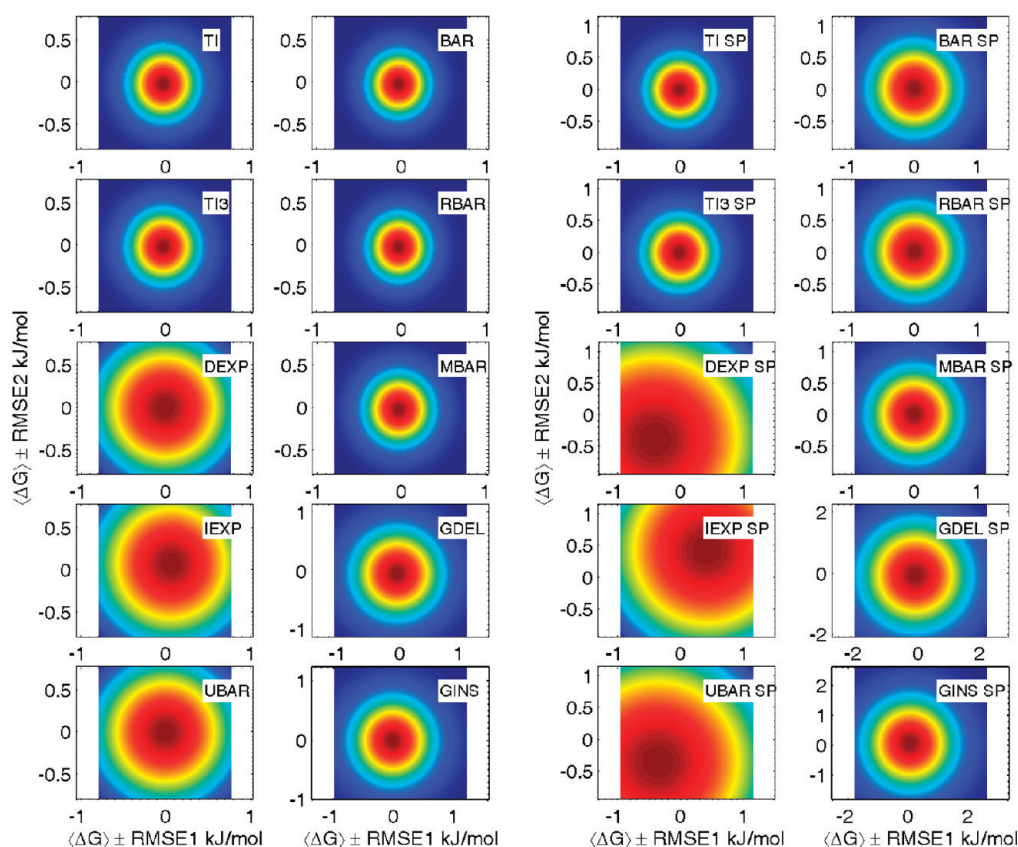


Figure 9. Bivariate Gaussian plots for dipole inversion. TI is reliable for this molecule, and DEXP, IEXP, GDEL, and GINS again are the least accurate and precise.

estimating free energy. TI and TI3 have more bias, which becomes substantial for sparse sets. IEXP and DEXP and especially GDEL and GINS have even larger bias. For dipole inversion, DEXP and IEXP show the largest bias both with respect to results from large number of samples and large number of intermediate states. All other methods for dipole inversion, including GINS and GDEL, show almost equal biases within statistical error limits of zero. For anthracene hydration free energies BAR, UBAR, RBAR, and MBAR show moderate biases for the full λ set compared to TI and TI3, but these results are again likely to be noise (see Table S9 in the Supporting Information). DEXP and IEXP as usual show high biases.

5.3. Overall Reliability. The knowledge of bias and uncertainty can now be put together to analyze the reliability of a method in estimating the free energy, which in this case we will define as consistently lowest RMSE. The bivariate Gaussian plots (Figures 8–10) show the reliability of a method, by visualizing the RMSE using both bias estimates for each test case. For each figure, the first two columns contain bivariate RMSE plots for full λ state runs and the last two show the results of the sparse λ state runs.

Figure 8, with data for UA methane solvation, shows that MBAR, RBAR, BAR, and UBAR have small and equal spreads in both horizontal and vertical directions indicating that these are reliable estimates of free energy both for sparse as well as full λ sets. TI and TI3 give reliable estimates of free energy only in full λ state runs but are dominated by bias due to number of intermediate states with the sparse set. IEXP has lower RMSE compared to TI and TI3. GDEL and GINS are unreliable in both the full and sparse λ sets.

In Figure 9 for dipole inversion, free energy estimates from TI and TI3 are comparable in reliability with MBAR, BAR, RBAR, and UBAR. GDEL and GINS also give fairly accurate estimates of free energy for dipole inversion given the poor performance in other systems. This can be explained in the light of the work done by Hummer, Pratt, and Garcia on free energy of ionic hydration,⁷⁴ in which they found that the electrostatic potential energy distribution follows Gaussian behavior. Thus even for large changes in charges, Gaussian methods may still be a viable and useful method.

The anthracene solvation test set is a harder problem, and no method is as accurate as with the other two molecular cases, as seen in Figure 10. TI, TI3, IEXP, BAR, RBAR, UBAR, and MBAR perform equally well within noise for full λ set. In the sparse λ set, however, IEXP and UBAR become significantly worse than the other methods with TI being slightly worse. Improved performance of TI relative to acceptance ratio methods for the anthracene test set is because the sparse λ set for anthracene solvation case (four states between end points) is not as aggressive as the methane solvation case (only one state between end points). GDEL, GINS, and DEXP are unreliable estimators of anthracene hydration free energy for both the full and sparse λ sets, with both low accuracy and precision in their predicted free energy and uncertainty estimates.

5.4. Testing the Gaussian Distribution of Free Energies. Asymptotic error estimate methods assume normal distribution of error, as does the use of a standard deviation to describe the error distribution. It is important to test if this assumption is actually valid. Figure 11 demonstrates graphically the distribution

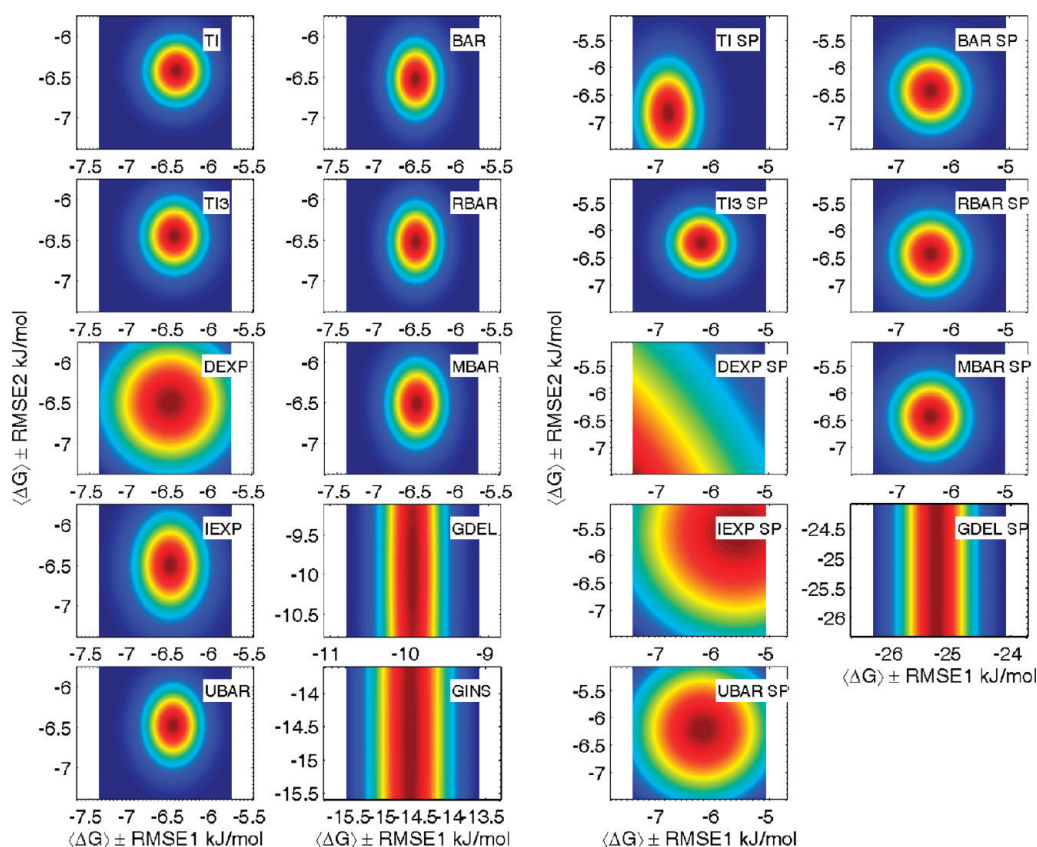


Figure 10. Bivariate Gaussian plots for anthracene solvation free energy. The effect of bias due to the number of intermediate states is evident in almost all methods as all spreads are elliptical in vertical direction. TI3 and MBAR both appear the most reliable, especially for the full λ set.

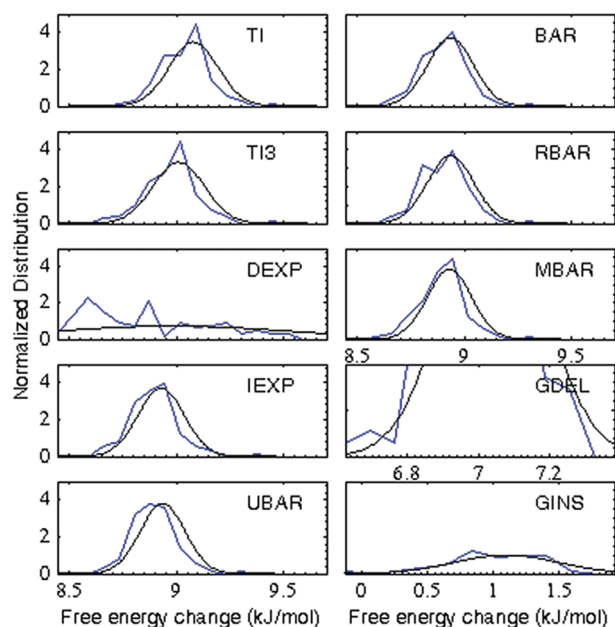


Figure 11. Subplots compare the distribution of estimated free energies from independent 100 repetitions (in blue) and Gaussian with the mean $\langle \Delta G \rangle$ and standard deviation $\langle \delta(\Delta G) \rangle$ from 100 independent repetitions (in black).

of free energy results for each test case, comparing it with a Gaussian with the mean and the variance of the distribution. The

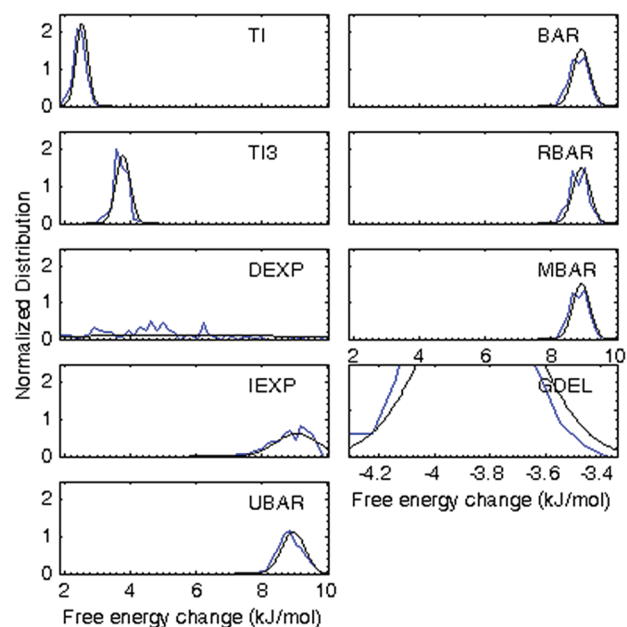


Figure 12. Each subplot compares the distribution of estimated free energies from 100 repetitions (in blue) and the Gaussian with the mean $\langle \Delta G \rangle$ and standard deviation $\langle \delta(\Delta G) \rangle$ from 100 repetitions (in black).

free energy distributions are plotted as histograms with the optimal bin width calculated from Scotts' formula,⁷⁵ with optimum bin width $h = 3.5 \sigma/n^{(1/3)}$, where n is the number of samples

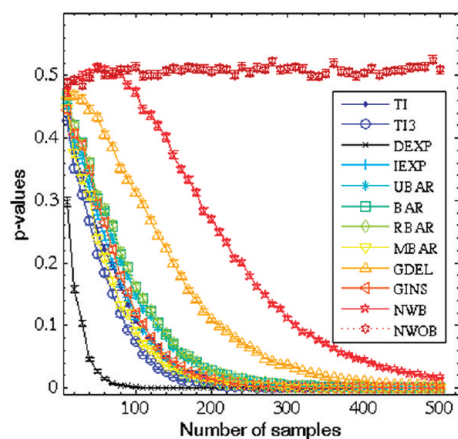


Figure 13. For methane solvation with full λ set, all p -value curves except DEXP are above 0.05 for 100 samples, resulting in the rejection of the null (normal distribution) hypothesis at the 95% confidence level for DEXP for as low as 40 samples.

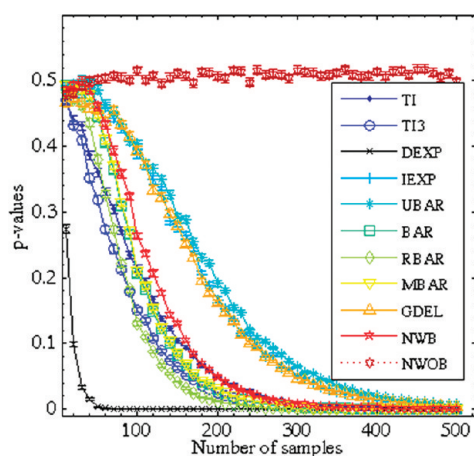


Figure 14. Methane solvation with the sparse λ set appears very similar to the full λ set case, with DEXP failing normality tests at 95% confidence level for as low as 25 samples. GINS is omitted because of a nonphysical (10^{32}) variance.

(i.e., 100), and σ is the standard deviation. For calculating the optimum bin width for the first eight methods, we have used the mean of the standard deviations from the 100 repetitions predicted by MBAR, though the variance is within a factor of 2–5 for all methods and only affects the qualitative visualization of the data. For GDEL and GINS, their own mean of predicted uncertainties is used to calculate the optimum bin width because uncertainties estimated with GDEL and GINS are drastically different from that of MBAR.

In Figure 11, the blue curve is the histogram of free energies, and the black curve is the Gaussian with the mean and the standard deviation estimated by the specified method. The shapes of the blue curve will match the black curve within noise if the distribution of free energies is indeed Gaussian, with the specified variance.

In Figures 11 and 12 and S1–S4 in the Supporting Information, we see that the distributions of free energies from all methods except DEXP well approximate the Gaussian, even when the variances are large. However, in several cases (GDEL and GINS)

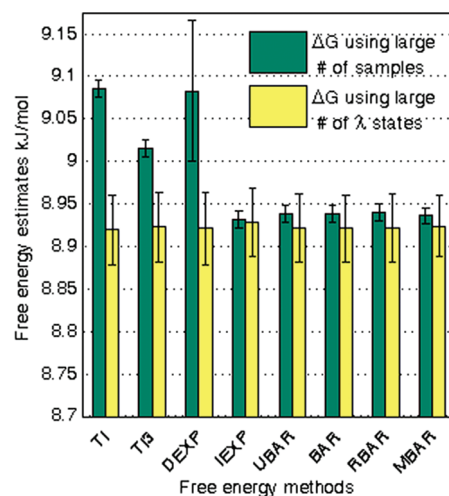


Figure 15. Free energies estimated using large number of intermediate states and a large number of samples converge to a single value for all free energy estimators.

the variances are so large that the comparisons become statistically meaningless. Besides DEXP, which fails in all cases IEXP and UBAR are sufficiently noisy for dipole inversion that the comparison to a Gaussian is problematic. For the sparse λ set IEXP and UBAR fail completely to be Gaussian. Similarly, for anthracene solvation IEXP and UBAR fail to be Gaussian for the sparse λ set. The 51 λ state results are omitted from GDEL and GINS plots as they lie outside the plot axis limits. Interestingly, we find that typically errors are distributed normally with the variance given by the analytical estimate even if the bias in the free energy is noticeable.

This visual analysis is only qualitative, and it is useful to also have quantitative tests of normality. We use the Shapiro–Wilk test for normality with an unknown mean and variance⁷³ to test the null hypothesis that a sample of free energies came from a normal population. The null hypothesis is rejected if the p -value is less than the specified probability. In this case, we choose a cutoff p -value of 0.05, corresponding to 95% confidence intervals. Importantly, the p -value depends on the number of samples; if a distribution is even the slightest different than a true Gaussian, with enough samples, the null hypothesis will always be rejected. Therefore for a continuous measure of normality, we calculate p -values as a function of number of samples and find the minimum number of samples on average that are required to reject the null hypothesis.

We randomly pick N ($10 < N < 500$) samples 2000 times with replacement from the original ensemble of 100 free energies, creating 2000 bootstrap replicas. For each bootstrap replica, we calculate the p -value. We then calculate the mean of the p -value over all the bootstrap replicas. This mean serves as an estimate of the p -value corresponding to N completely independent samples. Errors in the p -value are computed as the standard deviation over the 2000 bootstrap replicas. We use two separate controls for each method for this experiment. First, we generate N samples 2000 times from the Gaussian distribution with mean and variance from MBAR and calculate the corresponding p -values. Then to test the effect of bootstrapping from a finite distribution, we generate 100 normally distributed free energies, and then from this distribution, we randomly pick N samples 2000 times to calculate bootstrap replicas and again calculate p -values estimated as a

function of the number of samples. The control experiment using samples from a normal distribution with no bootstrap is abbreviated as NWOB, and the one which uses bootstrap is abbreviated as NWB.

Figures 13 and 14 and S5–S8 in the Supporting Information show average p -values as a function of number of samples pulled from the distribution for the 10 methods and two control cases for the three test systems. NWOB has the expected p -value = 0.5 curve, as the distribution is exactly Gaussian. NWB and all 10 methods have a decaying p -value curve with increasing numbers of samples. Decreasing p -value indicates increasing deviation from normality and increasing probability of rejecting the normality hypothesis. For NWB, although samples come from a known normal distribution, the null hypothesis is rejected at 95% confidence level when the number of samples is between 110 and 270 samples depending on the test set, indicating an approximate upper level of normality that can be observed with a finite bootstrapped set.

Examining Figures 13 and 14 and S5–S8 in the Supporting Information, we see that in almost all cases, for all methods other than DEXP, we would generally need more than 90 samples to reject the hypothesis that the samples were normal at a 95% confidence level, which should be sufficient for all general purposes. In all but the dipole inversion case, the null hypothesis of normality is rejected at the 95% confidence level for DEXP as low as 40 samples, which can be seen qualitatively in the distribution graphs in Figures 11 and 12 and S1–S4 in the Supporting Information. These results indicate for all methods except for DEXP, there are essentially no long tails or “black swans”, and the error estimates can be assumed to be Gaussians for all practical purposes. We note that variances that are known to be too large might require more samples to demonstrate that they are definitively not normal, but the accurate prediction of the magnitude of the variance is not measured by this statistical test, and the effects show up only in the large (more than 100) sample regime.

5.5. Convergence Properties. The true free energy estimate is not necessarily the experimental value of the free energy change of the process, but instead the infinite sampling limit of the particular choice of molecular model. We see that with a large number of intermediate states, all methods converge to the same value (Figure 15), whereas the 450 ns results with the sparse λ data set vary for different methods, with as usual large deviations seen in GDEL and GINS, indicating significant bias with respect to the overlap between states. Given sufficient sampling, increasing the number of λ states appears to be the best way to obtain asymptotic convergence to the true answer for most molecular transformations that have converged sampling.

5.6. Amount of Time Required for Free Energy Estimation Methods. We chose the calculation of anthracene solvation over 4.5 ns with the full λ set as our test system to compare computational time required by methods to compute free energies, because the system has the largest number of intermediate states with large free energy changes between states. We report the time required to calculate the free energies 201 times (for the original set and for 200 bootstrap sets) to eliminate variability caused by computational overhead in single calculations. The time required to read in data, make bootstrap samples, and perform book-keeping required by all methods was 249.5 s was subtracted from the total time to yield the analysis time of each method. Time, i.e., time required for generating samples is also not included in the analysis, as it is same for all methods.

Table 6. Analysis Time for Different Free Energy Methods^a

method	time taken (s)
TI	0.2
TI3	5.8
DEXP	13.5
IEXP	11.5
UBAR	15.3
BAR	93.5
RBAR	1148.2
MBAR	4913.5
GDEL	4.0
GINS	4.3

^aTimes are for 201 repetitions of a 4.5 ns dataset for anthracene solvation with 15 states. Time to generate samples is not included.

Table 7. Summary of All Statistical Tests

Reliability (accuracy plus bias) of free energy estimate (high to low)
MBAR > BAR = RBAR > UBAR > TI3 > TI > IEXP > DEXP > GDEL = GINS
Reliability of uncertainty estimate (high to low)
MBAR > TI3 = TI > BAR = RBAR > UBAR > IEXP > DEXP > GDEL = GINS
Computational cost of analysis of data (high to low)
MBAR > RBAR > BAR > UBAR > IEXP = DEXP > GDEL = GINS > TI3 > TI
Is distribution of estimated free energies Gaussian?
Yes for [MBAR, RBAR, BAR, UBAR, TI, TI3, IEXP, GDEL, GINS].
No for [DEXP]
Is bootstrap better than analytical uncertainty estimate?
Yes for [RBAR, BAR, UBAR, GDEL, GINS].
Both equally good for [MBAR, IEXP, DEXP, TI, TI3]

From Table 6 it is evident that MBAR takes the longest of all methods as it processes information from all the intermediate states to give an estimate of free energy and uncertainty. RBAR is the next most computationally costly. RBAR takes more time compared to BAR because multiple BAR calculations are performed over a range of free energies at each intermediate stage if a large range of possible values for the self-consistent constants are evaluated. UBAR takes less time compared to BAR because only a single iteration is performed at each stage. DEXP and IEXP are similar in cost UBAR. GDEL and GINS take less time compared to DEXP and IEXP. TI3 takes slightly longer compared to TI as it fits a spline of higher degree, but both are much cheaper than any others. However, the total time required even by MBAR is orders of magnitude less than the time required to perform the sampling, so the higher cost of MBAR is not an obstacle in most cases. Note that the time includes bootstrapping of 200 samples, so in the case of MBAR, where analytical variance is sufficient, the single calculation time is only 24 seconds.

6. CONCLUSIONS

In this paper we have proposed the first iteration of a set of test sets which can be used for benchmarking free energy calculation methods for small molecule solvation. We have demonstrated the utility of this test set by comparing 10 equilibrium free energy methods on 3 test cases for molecular solvation, with different spacing between intermediate states. We estimated the uncertainty in three different ways: the sample standard deviation, the

analytical estimate, and the bootstrap estimate as well as the uncertainty in each of these estimates of the uncertainty. We also calculated biases in free energy estimates at the large number of samples limit and large number of intermediate states limit separately. We graphically demonstrated the effect of the variance and two separate types of bias by bivariate Gaussian plots expressing the overall reliability of the methods. We demonstrated that bootstrap sampling accurately predicts the properties of the sample distribution observed from 100 independent simulations for all the free energy methods. We find that all uncertainty estimates for all but the worst performing methods are highly consistent, with relative errors of only 5% to 10%, and thus can be generally expected to be consistent from sample to sample. We also showed that the histogram of free energies from 100 independent simulations has a Gaussian form for TI, TI3, BAR, RBAR, UBAR, MBAR, GDEL, and GINS, but that DEXP and sometimes IEXP deviate from Gaussian distributions.

We have found that MBAR is the most reliable of all free energy estimators, showing consistency in accuracy and precision in both free energy and uncertainty prediction. TI and TI3 are better uncertainty estimators compared to BAR, UBAR, and RBAR, with equal performance to MBAR when sufficient intermediate states are included but are biased with respect to the number of intermediate states if such care is not taken. It is likely that this bias can be reduced with the judicious use of application-specific integration schemes, though the variance will not be reduced, since the uncertainties for the $\partial U/\partial\lambda$ observables do not change in alternative choices of the integration weights. When the $\partial U/\partial\lambda$ vs λ curve has low curvature, such as in dipole inversion, both TI and TI3 are equally reliable. But when the curve is nonlinear, i.e., when LJ spheres grow or disappear, TI3 gives better estimates of free energy than TI. BAR and RBAR have relatively negligible bias, but their uncertainty estimates are frequently underestimated by 25% to 30% when overlap between states is not negligible. UBAR is often as good as BAR and RBAR but can fail with low numbers of intermediate states. IEXP and DEXP are less reliable than TI and acceptance ratio methods and should be avoided if samples can be collected from all intermediates in both the forward or back direction or if the derivative of the Hamiltonian along the pathway can be computed. IEXP does work in some cases, but in general, IEXP and DEXP give poor estimates for uncertainty and free energies. GINS and GDEL do not compare well with the other methods in all the test systems except dipole inversion test case. They only work if there is large number of intermediate states or if the distributions are inherently Gaussian. However, even here they are not as accurate or precise as the other methods.

MBAR is the most expensive, but the amount of time required for analysis is orders of magnitude less than the time required for collecting data. UBAR takes less time compared to RBAR and BAR and should only be considered as a quick and easy (but not so reliable) alternative to BAR and RBAR. RBAR requires some knowledge of the maximum free energy gap but does not require storing all the energy data. BAR, like MBAR, requires storing energy data for later analysis but requires no knowledge about the size of the maximum free energy difference. GDEL and GINS take less time compared to DEXP and IEXP but they heavily sacrifice accuracy and precision for speed in virtually all cases. Finally, TI takes least time to estimate free energies and uncertainties but a little extra computation (fitting cubic splines) in TI3 improves the accuracy in free energy estimate using TI. We summarize these conclusions in Table 7.

Availability of the Benchmark Data Set for Multiple Simulation Packages. The benchmark test set is available for distribution and use at <http://www.alchemistry.org>. It contains starting configurations and parameter files for all 100 uncorrelated starting configurations in three formats, corresponding to three different molecular dynamics packages, GROMACS (*.gro, *.top, and *.mdp), AMBER (*.inpcrd, *.prmtop, and *.in), and DESMOND (*.cms and *.cfg) as well as detailed instructions for the test set's use. To ensure that the parameter files are correctly constructed, we have calculated the single point energies of the 100 structures using the input files and the corresponding MD package. We invite users of other simulation packages to contact us in order to add energy comparisons to these simulation packages.

This is intended to be the first version of the benchmark test set. It is not comprehensive and will require further expansion to be more useful to a wider range of researchers. We hope that these data sets will be useful for other researchers to compare other free energy estimators, such as alternative TI schemes, nonequilibrium free energy methods, and other more novel methods.

The intention is for future versions of this benchmark test set to be developed in response to feedback and as resources become available. Future versions of the benchmark will ideally include model molecules with long correlation time internal motion. A model system for the transformation of bonded terms would also be useful, though these transformations may have sufficiently short correlation times to make the differences in efficiency less relevant. It is also sometimes useful to compute molecular potentials of mean force in place of molecular transformations. Variance is directly related to phase space overlap, so the patterns discovered for alchemical transformations should also hold for construction of PMF. However a test system to validate this hypothesis will be useful. Finally, it would be ideal to eventually test the efficiency of methods to calculate free energies of well-studied ligand-binding systems and increasingly tractable protein–ligand binding systems, such as T4 lysozyme. We look forward to collaborating with other researchers to further develop the benchmark set and will post results and comparisons with other methods and between different codes on the <http://alchemistry.org> Web site along with the benchmark set data.

APPENDIX: THERMODYNAMIC INTEGRATION USING CUBIC SPLINES

We fit the $\langle(\partial U(\lambda)/\partial\lambda)_{\lambda_i}\rangle$ vs λ_i curve piecewise to a series of cubic polynomials $S_i(\lambda)$:

$$S_i(\lambda) = a_i + b_i(\lambda - \lambda_i) + c_i(\lambda - \lambda_i)^2 + d_i(\lambda - \lambda_i)^3 \quad \forall 1 \leq i \leq K - 1 \quad (\text{A1})$$

Here K is the total number of intermediate states, creating $K - 1$ intervals for splining. Each spline of a given interval has its own set of coefficients a_i , b_i , c_i , and d_i , which can be computed by standard linear algebra methods using the conditions that define a natural cubic spline.

$$S_i(\lambda_i) = a_i = \langle(\partial U(\lambda)/\partial\lambda)_{\lambda_i}\rangle \quad \forall 1 \leq i \leq K - 1 \quad (\text{A2})$$

$$S_i(\lambda_{i+1}) = \langle(\partial U(\lambda)/\partial\lambda)_{\lambda_{i+1}}\rangle \quad \forall 1 \leq i \leq K - 1 \quad (\text{A3})$$

$$S'_i(\lambda_{i+1}) = S'_{i+1}(\lambda_{i+1}) \quad \forall 1 \leq i \leq K - 2 \quad (\text{A4})$$

$$S_i''(\lambda_{i+1}) = S_{i+1}''(\lambda_{i+1}) \quad \forall 1 \leq i \leq K-2 \quad (\text{A5})$$

$$S_1''(\lambda_1) = 0 \quad (\text{A6})$$

$$S_{K-1}''(\lambda_K) = 0 \quad (\text{A7})$$

When we integrate piece wise over all the intervals, we get the total free energy change:

$$\Delta G = \sum_{i=1}^{K-1} \int_i^{i+1} d\lambda S_i(\lambda) \quad (\text{A8})$$

$$\Delta G = \sum_{i=1}^{K-1} a_i(\lambda_{i+1} - \lambda_i) + \frac{b_i}{2}(\lambda_{i+1} - \lambda_i)^2 + \frac{c_i}{3}(\lambda_{i+1} - \lambda_i)^3 + \frac{d_i}{4}(\lambda_{i+1} - \lambda_i)^4 \quad (\text{A9})$$

We need to write ΔG in the form described in eq 3, as a weighted sum of $\langle(\partial U(\lambda)/\partial \lambda)_{\lambda_i}\rangle$ at each λ , so that we can propagate the error using eq 4. We must solve for the coefficients a_i , b_i , c_i , and d_i in such a way such that they be expressed as a linear weighted sum of individual $\langle(\partial U(\lambda)/\partial \lambda)_{\lambda_i}\rangle$.

$$\begin{bmatrix} a_1 \\ \vdots \\ a_{K-1} \end{bmatrix} = \begin{bmatrix} A_{1,1} & \cdots & A_{1,K} \\ \vdots & \ddots & \vdots \\ A_{K-1,1} & \cdots & A_{K-1,K} \end{bmatrix} \begin{bmatrix} \langle(\partial U(\lambda)/\partial \lambda)_1\rangle \\ \vdots \\ \langle(\partial U(\lambda)/\partial \lambda)_K\rangle \end{bmatrix} \quad (\text{A10})$$

Here A_{ij} are the weights in a weight matrix for a_i . Similarly b_i , c_i , and d_i are expressed as linear weighted sums of $\langle(\partial U(\lambda)/\partial \lambda)_{\lambda_i}\rangle$ with B_{ij} , C_{ij} , and D_{ij} as the weights in the respective $K-1 \times K$ matrices. There exist a unique solution for a , b , c , and d . The A , B , C , and D matrices are all of rank $K-1$ and are invertible.

We can then finally combine these into a single weight matrix. Defining $h_i = \lambda_{i+1} - \lambda_i$, then eq A9 can be written as a linear weighted sum:

$$\Delta G = \sum_{i=1}^{K-1} \sum_{j=1}^K \left(h_i A_{ij} + \frac{h_i^2}{2} B_{ij} + \frac{h_i^3}{3} C_{ij} + \frac{h_i^4}{4} D_{ij} \right) \langle(\partial U(\lambda)/\partial \lambda)_j\rangle \quad (\text{A11})$$

Equation A11 can be further written as:

$$\Delta G = \sum_{i=1}^{K-1} \sum_{j=1}^K (W_{ij}) \langle(\partial U(\lambda)/\partial \lambda)_j\rangle \quad (\text{A12})$$

Here

$$W_{ij} = h_i A_{ij} + \frac{h_i^2}{2} B_{ij} + \frac{h_i^3}{3} C_{ij} + \frac{h_i^4}{4} D_{ij} \quad (\text{A13})$$

Once we have the weights, we can calculate the overall free energy change using eq A12 and the uncertainty estimate using the following equation.

$$\sigma_{\Delta G}^2 = \sum_{i=1}^{K-1} \sum_{j=1}^K W_{ij}^2 \sigma_j^2 \quad (\text{A14})$$

An implementation of this weighting for TI using GROMACS is included in the examples section of the pymbar distribution at <https://simtk.org/home/pymbar>.

■ ASSOCIATED CONTENT

S Supporting Information. Figures S1–S8 and Tables S1–S10 referred to in the text. It also contains the current version of files described in the section “Availability of the benchmark data set for multiple simulation packages.” The most recent version of the benchmark can be found at www.alchemistry.org. This material is available free of charge via the Internet at <http://pubs.acs.org>.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: michael.shirts@virginia.edu.

■ ACKNOWLEDGMENT

The authors acknowledge support for the Oak Ridge Associated Universities Ralph E. Powe, Jr. Faculty Enhancement Award. We also would like to acknowledge significant help from James Watney at D. E. Shaw Research for help in running Desmond.

■ REFERENCES

- (1) Davies, J. W.; Glick, M.; Jenkins, J. L. *Curr. Opin. Chem. Biol.* **2006**, *10*, 343–351.
- (2) Merz, K. M.; Ringe, D.; Reynolds, C. H. *Drug Design: Structure- and Ligand-Based Approaches*; Cambridge University Press: New York, 2010.
- (3) Mobley, D.; Dill, K. *Structure* **2009**, *17*, 489–498.
- (4) Bai, H.; Lai, L. *Acta Physicochim. Sin.* **2010**, *26*, 1988–1997.
- (5) Deng, Y.; Roux, B. *J. Phys. Chem. B* **2009**, *113*, 2234–2246.
- (6) Huang, N.; Kalyanaraman, C.; Bernacki, K.; Jacobson, M. P. *Phys. Chem. Chem. Phys.* **2006**, *8*, 5166–5177.
- (7) Zamolo, L.; Salvalaglio, M.; Cavallotti, C.; Galarza, B.; Sadler, C.; Williams, S.; Hofer, S.; Horak, J.; Lindner, W. *J. Phys. Chem. B* **2010**, *114*, 9367–9380.
- (8) Duren, T.; Bae, Y.; Snurr, R. *Chem. Soc. Rev.* **2009**, *38*, 1237–1247.
- (9) Nel, A. E.; Madler, L.; Velegol, D.; Xia, T.; Hoek, E. M. V.; Somasundaran, P.; Klaessig, F.; Castranova, V.; Thompson, M. *Nat. Mater.* **2009**, *8*, 543–557.
- (10) Liong, M.; Lu, J.; Kovochich, M.; Xia, T.; Ruehm, S. G.; Nel, A. E.; Tamanoi, F.; Zink, J. I. *ACS Nano* **2008**, *2*, 889–896.
- (11) Shi, X.; Wang, S.; Shen, M.; Antwerp, M.; Chen, X.; Li, C.; Petersen, E.; Huang, Q.; Weber, W.; Baker, J. *Biomacromolecules* **2009**, *10*, 1744–1750.
- (12) Nishide, H. *Advanced Nanomaterials*; Wiley-VCH: Hoboken, NJ, 2010.
- (13) Yuan, H.; Zhang, S. *Appl. Phys. Lett.* **2010**, *96*, 033704.
- (14) Fisher, E. H.; Rhodes, N. *Proc. Inst. Mech. Eng., Part C* **1996**, *210*, 91–94.
- (15) Roache, P. J. *Annu. Rev. Fluid Mech.* **1997**, *29*, 123–160.
- (16) Oberkampf, W. L.; Trucano, T. G. *Prog. Aerosp. Sci.* **2002**, *38*, 209–272.
- (17) Johnson, F.; Tinoco, E.; Yu, N. *Comput. Fluids.* **2005**, *34*, 1115–1151.
- (18) Kellar, W.; Savill, A.; Dawes, W. *Proc. High Perform. Comput. Networking* **1999**, *1593*, 90–98.
- (19) del Álamo, J. C.; Marsden, A. L.; Lasherasa, J. C. *Rev. Esp. Cardiol.* **2009**, *62*, 781–805.
- (20) Botar, C. C.; Vasile, T.; Sfrangeu, S.; Clichici, S.; Agachi, P. S.; Badea, R.; Mircea, P.; Cristea, M. V. In *20th European Symposium on Computer Aided Process Engineering*; Elsevier: Waltham, MA, 2010; Vol. 28, pp 205–210.
- (21) Sampson, B. *Prof. Eng.* **2007**, *20*, 37–37.
- (22) Oberkampf, W.; Trucano, T. *Nucl. Eng. Des.* **2008**, *238*, 716–743.

- (23) Johnson, R. D., III CCCBDB Computational Chemistry Comparison and Benchmark Database, *NIST Standard Reference Database Number 101*, release 15b, August 2011. <http://cccbdb.nist.gov/>.
- (24) Maginn, E. J.; Elliott, J. R. *Ind. Eng. Chem. Res.* **2010**, *49*, 3059–3078.
- (25) Kirkwood, J. G. *J. Chem. Phys.* **1935**, *3*, 300.
- (26) Bruckner, S.; Boresch, S. *J. Comput. Chem.* **2011**, *32*, 1320–1333.
- (27) Zwanzig, R. W. *J. Chem. Phys.* **1955**, *23*, 1915.
- (28) Shirts, M. R.; Pande, V. S. *J. Chem. Phys.* **2005**, *122*, 144107.
- (29) Bennett, C. H. *J. Comput. Phys.* **1976**, *22*, 245–268.
- (30) Shirts, M. R.; Chodera, J. D. *J. Chem. Phys.* **2008**, *129*, 124105.
- (31) Ytreberg, F. M.; Swendsen, R. H.; Zuckerman, D. M. *J. Chem. Phys.* **2006**, *125*, 184114.
- (32) These numbers were generated by adding the primary citations for each of the listed methods (23–30) and searches for thermodynamic integration (TI) and free energy perturbation in ISI Web of Science, as the original papers for these methods are older than the ISI database.
- (33) Ytreberg, F. M.; Zuckerman, D. M. *J. Chem. Phys.* **2004**, *120*, 10876.
- (34) Lin, C.; Wood, R. H. *J. Comput. Chem.* **1994**, *15*, 149–154.
- (35) Radmer, R.; Kollman, P. J. *J. Comput. Chem.* **1997**, *18*, 902–919.
- (36) Lelievre, T.; Stoltz, G.; Rousset, M. *Free Energy Computations: A Mathematical Perspective*; 1st ed.; Imperial College Press: London, 2010.
- (37) Gouda, H.; Kuntz, I. D.; Case, D. A.; Kollman, P. A. *Biopolymers* **2003**, *68*, 16–34.
- (38) Meirovitch, H.; Chelvaraja, S.; White, R. P. *Curr. Protein Pept. Sci.* **2009**, *10*, 229–243.
- (39) Jiang, W.; Roux, B. *J. Chem. Theory Comput.* **2010**, *6*, 2559–2565.
- (40) Zheng, L.; Chen, M.; Yang, W. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105*, 20227–20232.
- (41) Cencek, W.; Szalewicz, K. *Int. J. Quantum Chem.* **2008**, *108*, 2191–2198.
- (42) Gurkan, B.; Goodrich, B.; Mindrup, E.; Ficke, L.; Massel, M.; Seo, S.; Senftle, T.; Wu, H.; Glaser, M.; Shah, J.; Maginn, E.; Brennecke, J.; Schneider, W. *J. Phys. Chem. Lett.* **2010**, *1*, 3494–3499.
- (43) Oden, J. T.; Belytschko, T.; Fish, J.; Hughes, T. J. R.; Johnson, C.; Keyes, D.; Laub, A.; Petzold, L.; Srolovitz, D.; Yip, S. *A Report of the National Science Foundation Blue Ribbon Panel on Simulation-Based Engineering Science: Revolutionizing Engineering Science through Simulation*; National Science Foundation: Arlington, VA, 2006.
- (44) Glotzer, S.; Kim, S.; Cummings, P.; Deshmukh, A.; Head-Gordon, M.; Karniadakis, G.; Petzold, L.; Sagui, C.; Shinozuka, M. *WTEC Panel Report on International Assessment of Research and Development in Simulation-Based Engineering and Science*; World Technology Evaluation Center, Inc.: Baltimore, Maryland, 2009.
- (45) Pohorille, A.; Jarzynski, C.; Chipot, C. *J. Phys. Chem. B* **2010**, *1420*–1426.
- (46) Radmer, R.; Kollman, P. J. *J. Comput. Chem.* **1997**, *18*, 902–919.
- (47) Jorge, M.; Garrido, N. M.; Queimada, A. J.; Economou, I. G.; Macedo, E. A. *J. Chem. Theory Comput.* **2010**, *6*, 1018–1027.
- (48) Kofke, D. A.; Cummings, P. T. *Mol. Phys.* **1997**, *92*, 973–996.
- (49) Hummer, G. *J. Chem. Phys.* **2001**, *114*, 7330.
- (50) Efron, B.; Tibshirani, R. J. *An Introduction to the Bootstrap*; 1st ed.; Chapman and Hall/CRC: Norman, OK, 1994; Vol. 5.
- (51) Jorgensen, W. L.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1988**, *110*, 1657–1666.
- (52) Hünenberger, P. H.; McCammon, J. A. *J. Chem. Phys.* **1999**, *110*, 1856.
- (53) Pitera, J.; Van Gunsteren, W. *Mol. Simulat.* **2002**, *28*, 45–65.
- (54) Martínez-Veracoechea, F. J.; Escobedo, F. A. *J. Phys. Chem. B* **2008**, *112*, 8120–8128.
- (55) Wang, F.; Landau, D. P. *Phys. Rev. Lett.* **2001**, *86*, 2050.
- (56) Jarzynski, C. *Phys. Rev. E* **1997**, *56*, 5018.
- (57) Resat, H.; Mezei, M. *J. Chem. Phys.* **1993**, *99*, 6052.
- (58) Mark, A. E.; van Helden, S. P.; Smith, P. E.; Janssen, L. H. M.; van Gunsteren, W. F. *J. Am. Chem. Soc.* **2011**, *116*, 6293–6302.
- (59) Wu, D.; Kofke, D. A. *J. Chem. Phys.* **2005**, *123*, 054103.
- (60) Barker, A. *Aust. J. Phys.* **1965**, *18*, 119–134.
- (61) Fenwick, M. K.; Escobedo, F. A. *J. Chem. Phys.* **2004**, *120*, 3066.
- (62) Ferrenberg, A. M.; Swendsen, R. H. *Phys. Rev. Lett.* **1989**, *63*, 1195.
- (63) Ferrenberg, A. M.; Swendsen, R. H. *Phys. Rev. Lett.* **1988**, *61*, 2635.
- (64) Souaille, M.; Roux, B. *Comput. Phys. Commun.* **2001**, *135*, 40–57.
- (65) Schuttelkopf, A.; van Aalten, D. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2004**, *60*, 1355–1363.
- (66) Hawkins, P. C. D.; Skillman, A. G.; Warren, G. L.; Ellingson, B. A.; Stahl, M. T. *J. Chem. Inf. Model.* **2010**, *50*, 572–584.
- (67) Xiao, Y.; Li, D. *Math. Meth. Oper. Res.* **2008**, *67*, 443–454.
- (68) Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids*, new ed.; Clarendon Press: Oxford, U.K., 1989.
- (69) Wu, X.; Brooks, B. R. *Chem. Phys. Lett.* **2003**, *381*, 512–518.
- (70) Hess, B. *J. Chem. Phys.* **2002**, *116*, 209.
- (71) Martyna, G. J.; Tuckerman, M. E.; Tobias, D. J.; Klein, M. L. *Mol. Phys.* **1996**, *87*, 1117.
- (72) Kofke, D. *Mol. Phys.* **2006**, *104*, 3701–3708.
- (73) Shapiro, S. S.; Wilk, M. B. *Biometrika* **1965**, *52*, 591–611.
- (74) Hummer, G.; Pratt, L. R.; García, A. E. *J. Phys. Chem.* **1996**, *100*, 1206–1215.
- (75) Scott, D. W. *Biometrika* **1979**, *66*, 605–610.


Micellization Studied by GPU-Accelerated Coarse-Grained Molecular Dynamics

Benjamin G. Levine,^{†,∇} David N. LeBard,^{†,∇} Russell DeVane,[‡] Wataru Shinoda,[§] Axel Kohlmeyer,[†] and Michael L. Klein^{*,†}

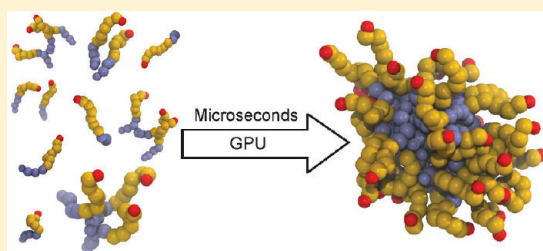
[†]Institute for Computational Molecular Science and Department of Chemistry, Temple University, Philadelphia, Pennsylvania 19122, United States

[‡]The Procter & Gamble Company, Cincinnati, Ohio 45253, United States

[§]National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Ibaraki, Japan

 Supporting Information

ABSTRACT: The computational design of advanced materials based on surfactant self-assembly without ever stepping foot in the laboratory is an important goal, but there are significant barriers to this approach, because of the limited spatial and temporal scales accessible by computer simulations. In this paper, we report our work to bridge the gap between laboratory and computational time scales by implementing the coarse-grained (CG) force field previously reported by Shinoda et al. [Shinoda, W.; DeVane, R.; Klein, M. L. *Mol. Simul.* **2007**, *33*, 27–36] into the HOOMD-Blue graphical processing unit (GPU)-accelerated molecular dynamics (MD) software package previously reported by Anderson et al. [Anderson, J. A.; Lorenz, C. D.; Travesset, A. J. *Comput. Phys.* **2008**, *227*, 5342–5359]. For a system of 25 750 particles, this implementation provides performance on a single GPU, which is superior to that of a widely used parallel MD simulation code running on an optimally sized CPU-based cluster. Using our GPU setup, we have collected 0.6 ms of MD trajectory data for aqueous solutions of 7 different nonionic polyethylene glycol (PEG) surfactants, with most of the systems studied representing $\sim 1\,000\,000$ atoms. From this data, we calculated various properties as a function of the length of the hydrophobic tails and PEG head groups. Specifically, we determined critical micelle concentrations (CMCs), which are in good agreement with experimental data, and characterized the size and shape of micelles. However, even with the microsecond trajectories employed in this study, we observed that the micelles composed of relatively hydrophobic surfactants are continuing to grow at the end of our simulations. This suggests that the final micelle size distributions of these systems are strongly dependent on initial conditions and that either longer simulations or advanced sampling techniques are needed to properly sample their equilibrium distributions. Nonetheless, the combination of coarse-grained modeling and GPU acceleration marks a significant step toward the computational prediction of the thermodynamic properties of slowly evolving surfactant systems.



1. INTRODUCTION

Self-assembly, whereby solutions of amphiphilic molecules spontaneously order themselves into more-complex structures, is a rich and important physical phenomenon.^{3,4} In addition to playing a vital role in biology and in many industrial and consumer applications, self-assembly has been harnessed to create human-designed nanostructures and functional materials.^{5,6} Over the last 20 years, computer simulation has emerged as an important method to study self-assembly processes.^{7–14} Recent simulation studies have addressed the properties of micellar^{15–26} and reverse micellar^{27–29} solutions, as well as more-complex phases^{9,30–36} and dynamical behaviors.^{37,38} Molecular dynamics (MD) techniques have been applied to examine self-assembled macromolecular materials as well, such as those composed of multiblock copolymers³⁹ and tethered nanoparticles.⁴⁰

While much progress has been made, the quantitative prediction of the thermodynamic and kinetic properties of self-assembled solutions by computer simulation remains a challenging prospect.

System sizes and time scales of processes accessible by MD methods are limited by the current computational power available. Many advances in MD algorithms,^{41–43} sampling techniques,^{44,45} and computer hardware⁴⁶ are driven by the need to study larger systems over longer time scales. By intelligently combining these techniques to accelerate their simulations, researchers can increase their capability by orders of magnitude and do science that was not previously possible.

One approach to extending the size and time scales accessible in MD simulations is coarse graining (CG), which sacrifices the atomic resolution present in all-atom (AA) force fields in exchange for greater computational efficiency. In contrast to AA models, CG models represent molecules as being composed of fictional *beads*, each representing more than one atom. Several CG force fields have been introduced in recent years, which differ

Received: July 27, 2011

Published: October 28, 2011

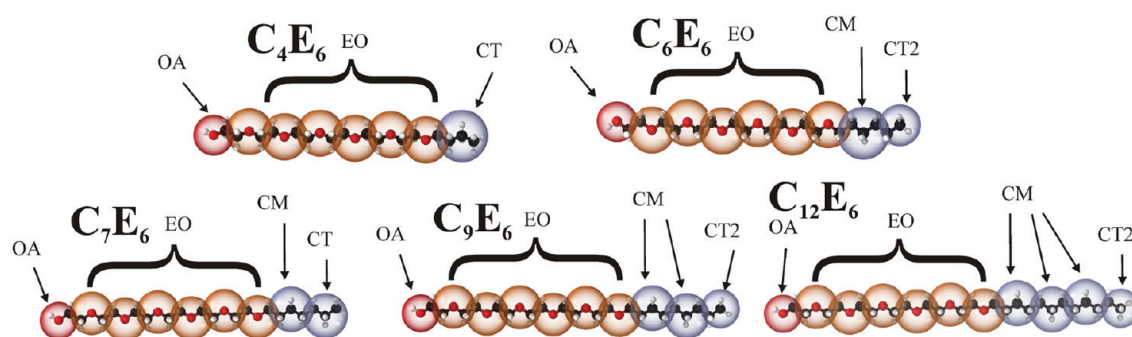


Figure 1. Illustrated representation of the SDK CG model, with CG beads overlaid on the underlying atomistic structure of a series of PEG surfactants considered in this work. Bead types are defined for terminal alcohols (OA), ethoxylate units (EO), nonterminal propyl groups (CM), terminal propyl groups (CT), and terminal ethyl groups (CT2).

in the degree of coarse graining (the number of atoms per bead) and the procedure by which they are parametrized.^{1,47,48}

In this work, we will study nonionic surfactant self-assembly with the CG force field of Shinoda, DeVane, and Klein (SDK).¹ This model has been shown to predict the self-assembly of micellar, lamellar, and hexagonal phases of nonionic polyethylene glycol (PEG) surfactant solutions in good agreement with experimental phase diagrams, despite the fact that no phase information was included in the parametrization.³⁵ The SDK approach to CG parametrization has been applied to many self-assembling systems including ionic surfactants (sulfates, sulfonates)³³ and zwitterionic lipids.³⁶

While coarse-grained models offer researchers one means of accelerating the discovery process, further acceleration is possible through the use of advanced computer hardware, such as graphical processing units (GPUs). GPUs are high-performance computer processors designed to accelerate graphical applications (e.g., computer games). In recent years, many scientists have begun to take advantage of the similarity between computer games and scientific simulation—specifically that the bottleneck of both is floating point math—using GPUs to perform scientific simulations.⁴⁹ In addition to their strong floating point performance, GPUs are appealing to scientists because of their relatively low cost, with current state-of-the-art models available for ~500 USD. Many researchers in computational chemistry are developing software that uses this advanced computer hardware,^{2,50–73} including the implementation of coarse-grained molecular modeling targeted at protein–protein interactions.⁷⁴ With two large-scale supercomputers based on GPU technology now available through the NSF Teragrid program^{75,76} and three large scale GPU-based machines listed among the four fastest supercomputers in the world,⁷⁷ it seems that GPU technology has grown from a promising new development into one of the pillars of high-performance computing.

One rapidly maturing implementation of GPU-based molecular dynamics is HOOMD-Blue.⁵⁰ Unlike many existing CPU-based codes that were modified to take advantage of GPU technology, HOOMD-Blue was written from the ground up to take advantage of the advanced features of NVidia GPUs, and it has been shown to provide performance comparable to that of a cluster of tens of CPU cores for medium-sized MD systems (tens to hundreds of thousands of particles).

In this paper, we present an efficient approach to studying self-assembly in micellar solutions using HOOMD-Blue in conjunction with the SDK CG MD force field. In section 2 of this paper,

we describe the CG force field employed in this work and our implementation of it into HOOMD-Blue. In section 3, we present data benchmark the performance of HOOMD-Blue in this context, report thermodynamic and structural properties calculated for a series of PEG surfactants, and analyze the convergence and accuracy of these results. In section 4, we draw conclusions.

2. METHODOLOGY

2.1. CG Potential for Aqueous PEG Surfactant Solutions.

Before discussing the details of the SDK model, the principles by which it was parametrized are presented. This model was developed to reproduce the thermodynamic and structural properties of complex soft materials.¹ Whenever possible, it has been parametrized to reproduce the experimentally determined thermodynamic properties of bulk materials (e.g., liquid water, hydrocarbons, ethers) and the interfaces between such materials. For example, the SDK model of liquid water was parametrized to faithfully represent the density and surface tension of liquid water under ambient conditions.⁷⁸ In addition to experimental data, structural information from AA simulations was also used as target data in the parametrization approach. Note that the SDK model was not parametrized with any knowledge of micellar PEG surfactant systems themselves; thus, all of the results reported in this work are predictions of the model.

The goal of the CG approach is to accurately describe materials without the need to treat the atoms of the system individually, as would be done in traditional AA simulations. As such, the materials are partitioned into CG beads, each of which represent, on average, three heavy atoms and associated hydrogens.^{1,35} The PEG surfactants studied here, which contain 49–81 atoms per monomer, are represented by flexible chains whose length vary from 7 to 11 CG beads (see Figure 1). Water normally accounts for the largest consumption of computational time in molecular simulations of aqueous systems; as such, its coarse graining is essential if a large savings of computer time is to be achieved. A previous study has determined that the properties of liquid water listed above can be accurately reproduced with three water molecules combined into a single CG bead.⁷⁸

The SDK potentials employed in this work to study the self-assembly of PEG surfactants involve three energy terms: angles, bonds, and nonbonding Lennard-Jones terms with nonstandard exponents.

Table 1. Simulated Systems

system	concentration (mM)	No. of surfactants	No. of water beads	No. of replicas	total sim. time (μ s)
C ₄ E ₆	129	700	96000	20	39.0
	247	1400	96000	4	3.4
	342	2000	96000	4	3.4
	440	2000	72000	4	4.4
	616	2000	48000	20	60.2
	771	2000	36000	4	7.3
	950	2000	24000	4	7.3
C ₆ E ₄	129	700	96000	20	70.9
C ₆ E ₆	128	700	96000	20	68.0
C ₆ E ₈	127	700	96000	20	69.8
C ₇ E ₆	128	700	96000	20	76.4
	65.5	350	96000	4	13.6
	30.3	160	96000	4	12.9
	15.3	80	96000	4	13.3
	7.7	40	96000	4	12.9
	3.8	20	96000	4	13.6
C ₉ E ₆	127	700	96000	20	85.8
C ₁₂ E ₆	126	700	96000	20	31.6
Total				200	593.8

The form of the nonbonded Lennard-Jones potentials will be collectively referred to as $U_{LJ}^{CG}(r)$:

$$U_{LJ}^{CG}(r) = \begin{cases} \frac{27}{4}\epsilon \left[\left(\frac{\sigma}{r}\right)^9 - \left(\frac{\sigma}{r}\right)^6 \right] & \text{for } U_{LJ}^{9-6} \\ \frac{3\sqrt{3}}{2}\epsilon \left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^4 \right] & \text{for } U_{LJ}^{12-4} \end{cases} \quad (1)$$

where r is the distance between a pair of beads, the minimum of the energy is given by ϵ , and σ is the distance where $U_{LJ}^{CG}(r) = 0$. Terms of the form U_{LJ}^{12-4} are used for all interactions involving water beads, while terms of the form U_{LJ}^{9-6} are used for interactions of all other pairs.¹

The total angle potential can be written as

$$U_{\text{ang}}^{CG}(\theta_{ijk}, r_{ik}) = U_{\text{ang}}(\theta_{ijk}) + U_{\text{soft}}(r_{ik}; r_c) \quad (2)$$

where i, j , and k index beads such that i and k are both bonded to j , and r_{ik} is the distance between the pair of particles i and k . U_{ang} is a standard harmonic angle potential, given by

$$U_{\text{ang}}(\theta_{ijk}) = \frac{K_{\text{ang}}}{2}(\theta_{ijk} - \theta_0)^2 \quad (3)$$

where θ_{ijk} is the angle between the ji and jk vectors, K_{ang} is the angular force constant, and θ_0 is the angle of minimum energy. An additional repulsive potential between particles i and k , $U_{\text{soft}}(r)$ is given by

$$U_{\text{soft}}(r; r_c) = \begin{cases} U_{LJ}^{CG}(r) - U_{LJ}^{CG}(r_c) & \text{for } r \leq r_c \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

where r_c is a cutoff defined by the Lennard-Jones parameters of beads i and k :

$$r_c = \begin{cases} \sigma \left(\frac{3}{2}\right)^{1/3} & \text{for } U_{LJ}^{9-6} \\ \sigma 3^{1/8} & \text{for } U_{LJ}^{12-4} \end{cases} \quad (5)$$

The use of this additional repulsive potential allows for a softer harmonic term while still maintaining a valid representation of excluded volume interactions between atoms i and k .

A simple harmonic bonding potential represents bonds between CG beads:

$$U_{\text{bond}}(r_{ij}) = \frac{K_{\text{bond}}}{2} (r_{ij} - r_0)^2 \quad (6)$$

where r_{ij} is the distance between beads i and j , r_0 is the minimum energy distance, and K_{bond} is the force constant. All simulation parameters used in this study can be found in the Supporting Information. Note that, because of the fact that harmonic potentials are implemented differently in HOOMD than in many other codes, the expressions for harmonic bond and angle potentials reported here differ from those reported previously¹ by a factor of 1/2. The reported parameters reflect this difference.

2.2. Potential Implementation. In order to run GPU-accelerated MD with the SDK potential, we chose to implement the energy terms associated with this force field into HOOMD-Blue,² which is a fast and easy-to-use implementation of MD designed to run on a single GPU. All potentials used in this study have been incorporated into the main HOOMD-Blue distributions, which are available for free download.⁷⁹

Harmonic bonding potentials were implemented in the original HOOMD-Blue distribution; therefore, we needed only to implement the CG angles and nonstandard Lennard-Jones potentials ($U_{LJ}^{9-6}, U_{LJ}^{12-4}$) from the SDK model.^{1,35} The GPU implementation for the CG force field potential is straightforward and follows the general framework for the short-ranged nonbonded forces provided in the original paper describing the implementation of HOOMD-Blue (see Algorithm 2 in ref 2). In HOOMD-Blue, a force kernel is launched that is comprised of several threads that run simultaneously. Each thread is associated with an individual particle of the system, for which it accumulates the forces. Positional data for the neighboring atoms is acquired from the fast texture memory of the GPU. In this arrangement, all force terms are calculated multiple times (Newton's third law is not used) to eliminate the need for communication between threads.²

2.3. Molecular Dynamics Simulations. To characterize the behavior of nonionic surfactant self-assembly, we have run simulations of aqueous solutions of various PEG surfactants. In total, 200 individual replica simulations were run, the details of which are listed in Table 1. All simulations presented here employed three-dimensional (3D) periodic boundary conditions. They were run in the NPT ensemble with a temperature of 303 K and a pressure of 1 atm, using a Nose-Hoover chain thermostat ($\tau = 100$ ps) and an isotropic Andersen barostat ($\tau_p = 100$ ps). All simulations were carried out using HOOMD-Blue.² Most were performed on Longhorn, a cluster of NVidia Quadro FX 5800 GPUs at the Texas Advanced Computing Center (TACC),⁷⁵ with the remainder being performed on local resources.

To investigate relationship between micellization behavior and hydrophobic tail length, we choose a series of five PEG surfactants for study: C₄E₆, C₆E₆, C₇E₆, C₉E₆, and C₁₂E₆. (Representative configurations are shown in Figure 2; throughout

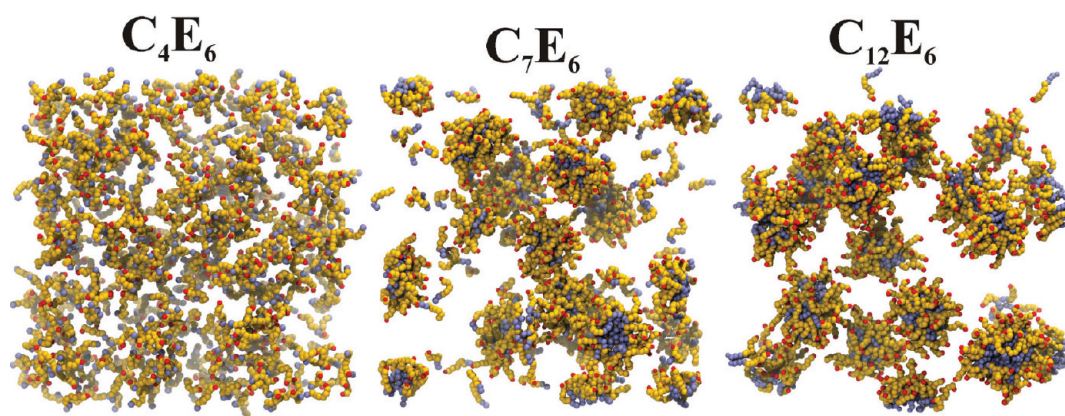


Figure 2. Representative configurations shown for a series of systems differing in hydrocarbon chain length. We illustrate the typical morphology of solutions ranging from nonmicellar (C_4E_6) to strongly micellar ($C_{12}E_6$) PEG surfactants. Water beads have been removed for clarity.

this paper, PEG surfactants will be abbreviated C_xE_y , which x is the length of the hydrophobic tail in carbon atoms and y is the length of the PEG chain in ethoxylate units.) In addition, we considered the dependence of behavior on the surfactant concentration by simulating a series of C_4E_6 and C_7E_6 solutions run at concentrations surrounding the critical micelle concentration (CMC). A study of the dependence of the CMC on the PEG chain length is carried out by considering a series of surfactants of the form C_6E_y . All systems presented here were started from random configurations, with no initial aggregation of the monomers. A time step of 10 fs was used, and trajectory data was saved every 200 ps for subsequent analysis. A total of 20 replicas started from different random initial conditions were run for most systems to improve statistical sampling. In total, this study presents ~ 0.6 ms of aggregate trajectory data.

2.4. Trajectory Analysis. From the trajectory data, aggregates were identified using a cutoff-cluster method, implemented in similar styles as in previous atomistic²⁴ and CG studies³⁵ of surfactant self-assembly. A cutoff of 8.2 Å was chosen for all trajectory data, because it represents the minimum in the radial distribution function of the CG hydrocarbons of micellar PEG from previous simulations using the same model.³⁵ An analysis of C_7E_6 shows only a subtle dependence of the micelle size distribution on the cutoff over the range between 6 Å and 12 Å (see the Supporting Information).

Because the simulations reported here begin from random, unaggregated arrangements of surfactants, it is necessary to exclude an initial equilibration period from thermodynamic averages. The length of this period depends on the rate of micelle formation and varies from 0.4 μ s to 2 μ s for the various systems.

In the course of the analysis presented in this paper, it is necessary to define *free monomers*. For the purposes of this paper, “free monomers” are defined to be any monomers that compose aggregates that are smaller than the first minimum in the micelle aggregation number distribution ($N_{\text{agg,min}}$; Table 2) for each specific surfactant. To illustrate, monomers composing aggregates smaller than $N_{\text{agg}} = 15$ are considered free monomers for C_6E_6 . This definition of “free monomer” includes both true free monomers (those that are completely solvated by water) and small premicellar aggregates, which are believed to exist in micellar solutions at concentrations near the CMC.^{80,81}

3. RESULTS AND DISCUSSION

3.1. GPU-Accelerated MD Performance. Before discussing the data collected using our GPU-accelerated MD approach,

Table 2. Properties of the Micelles Observed in Coarse-Grained (CG) Molecular Dynamics (MD) Simulations^a

system	conc (mM)	$N_{\text{agg,min}}$	$N_{\text{agg,max}}$	CMC (mole fract)	$R_{g,\text{max}}$ (Å)
C_6E_4	129	15	41	0.99×10^{-3}	15.75
C_6E_6	128	15	24	1.51×10^{-3}	15.75
C_7E_6	128	11	41	3.39×10^{-4}	17.75
C_9E_6	127	4	41	1.19×10^{-5}	18.25
$C_{12}E_6$	126	2	45	1.41×10^{-7}	18.25

^a The aggregation number at the first minimum of the micelle aggregation number distribution is labeled $N_{\text{agg,min}}$; $N_{\text{agg,max}}$ is the aggregation number at the maximum; and $R_{g,\text{max}}$ is the most probable value of the radius of gyration (R_g). Data for C_4E_6 and C_6E_8 is not included, because there is no clear division between micelles and smaller aggregates for these more-hydrophilic surfactants.

we report the performance of our implementation. For benchmarking purposes, a system representing a PEG surfactant solution one-quarter the size of our production simulations was chosen: 175 C_9E_6 molecules solvated in 24 000 CG water beads. The performance of HOOMD-Blue running on two different NVidia GPUs—a Quadro FX 5800, as is available on the TACC Longhorn cluster, and a GTX 480, which is a consumer graphics card based on the new Fermi GPU processor—is reported in Figure 3. For comparison, performance is also reported for the widely used and highly optimized CPU-based MD code LAMMPS^{82,83} running on the Cray XT5 supercomputer at the National Institute for Computational Science (Kraken). The performance on a single Quadro FX 5800 outperforms 64 processor cores of the XT5 running in parallel, while the newer GTX 480 hardware outperforms all configurations of CPUs tested.

These tests indicate the value of the GPU-acceleration strategy employed by HOOMD-Blue. HOOMD-Blue was designed from the ground up to run on a single GPU, thus eliminating the need for a significant amount of communication between the GPU and the main memory or CPU. While it is certainly true that massively parallel implementations of MD such as that in LAMMPS are ideal when simulating extremely large systems, for medium-sized systems, such as those simulated in this work (10 000–100 000 particles), a single GPU processor can give performance competitive with a much more expensive CPU cluster. If scientists proceed as if each GPU is its own small CPU

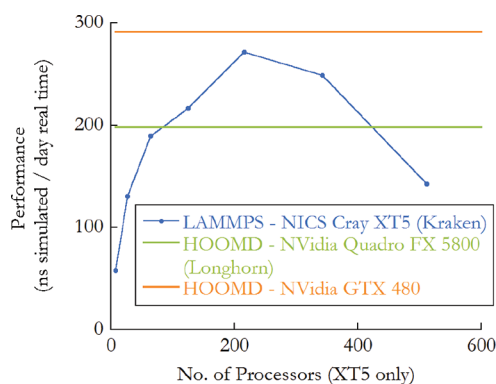


Figure 3. Performance of molecular dynamics (MD) simulations of a CG surfactant systems comprising 175 C_9E_6 molecules solvated in 24 000 CG water beads run using HOOMD-Blue on a single GPU is reported. Results are presented for two different NVidia GPUs: the Quadro FX 5800 (green) and GTX 480 (orange). The performance of LAMMPS, the highly optimized CPU-based MD code,^{82,83} running on the NICS Cray XT5 (Kraken) computer cluster is reported as a function of the number of processor cores employed for comparison (blue). The GPU implementation of the SDK model in HOOMD-Blue run on a single NVidia Quadro FX 5800 GPU outperforms LAMMPS running on 64 CPU cores. The same implementation running on a single Fermi-based NVidia GTX 480 card outperforms all configurations of CPUs tested.

cluster, it allows them to dramatically improve their sampling by running multiple replicas in parallel, and this is the strategy that we have employed in this work. Very intriguing, though, is the combination of LAMMPS' massively parallel computation strategy and GPU acceleration, which has shown great promise for extending the capabilities of computers to address questions regarding large-scale molecular systems.⁵⁴

3.2. Size and Structure of Micelles as a Function of Tail Length. From the 0.6 ms of trajectory data collected in this study, we have calculated a variety of thermodynamic properties of nonionic surfactant solutions as a function of the hydrophobic tail length, PEG chain length, and total concentration. A summary of the properties of many of the final micellar solutions can be found in Table 2.

First we investigate the dependence of the structural properties of micelles on the hydrophobic tail length by considering a series of surfactants of the form C_xE_6 . All systems were simulated with a mole fraction of monomer of 2.38×10^{-3} (700 surfactant molecules solvated in 96 000 CG water beads). The micelle aggregation number (N_{agg}) distributions of this series are shown in Figure 4. For C_4E_6 , the most hydrophilic surfactant considered, the probability of an aggregate decreases monotonically with its N_{agg} and no aggregates of >12 monomers are observed, indicating that either the concentrations of our simulations is below the CMC or that C_4E_6 does not form micelles. The properties of C_4E_6 will be examined in more detail below.

More hydrophobic surfactants, starting with C_6E_6 , exhibit a peak in the distribution corresponding to stable micelles. As has been observed in experiments,^{4,84} the peak of the micelle size distribution occurs at larger N_{agg} as the hydrophobic tail length increases from 6 to 7. Surprisingly, however, the maximum of the distribution does not shift to significantly larger N_{agg} as the hydrophobic tail length increases from 7 to 12, despite the fact that experiments^{4,84} demonstrate that it does. This discrepancy stems from the fact that the dynamics of micelle growth in these systems are very slow,

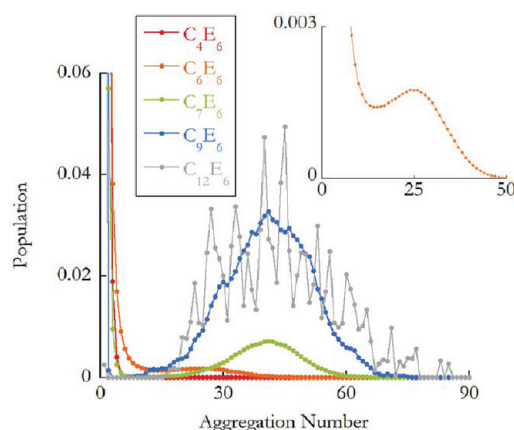


Figure 4. Distributions of micelle aggregation numbers (N_{agg}) of five different surfactant systems at the same concentration. Data are obtained from MD trajectories after a significant equilibration period (0.4–2 μ s, depending on the rate of micelle formation). For clarity, the inset shows the C_6E_6 distribution over a narrower range of aggregation number and population. No maximum in the distribution corresponding to stable micelles occurs for C_4E_6 , but more micelles form with increasing hydrophobic chain length. The aggregation number at which the distribution is at a maximum increases from 24 for C_6E_6 to 41 for C_7E_6 , but then stays relatively constant as the hydrophobic chain length increases to 9 and 12. As can be seen in Figure 5, this is likely due to the fact that these more hydrophobic surfactants would require a longer time to reach equilibrium. The probability of micelles composed of more than 60 monomers rises as the chain length increases from 7 to 12, despite the fact that the maximum of the distribution remains stationary.

compared to the time scales accessed by the present simulations. The convergence of the micelle size distributions will be investigated in more detail in the next subsection. The noise observed in the $C_{12}E_6$ is the result of the shorter sampling time and slower evolution of these simulations, compared with the more hydrophilic systems.

Though quantitative information about the micelle size distributions cannot be extracted from our MD data for these more hydrophobic surfactants, one can confirm that the proper qualitative trend—the average micelle aggregation number increases with increasing hydrophobic tail length—is reproduced by considering the time dependence of the average micelle number (Figure 5). Although the micelles observed for surfactants with a hydrophobic tail longer than six carbons continue to grow for the duration of the simulations, the more hydrophilic systems are converging toward lower values of N_{agg} compared to those that are more hydrophobic.

The shape of the simulated micelles is described in Figure 6, which shows histograms of micelle population as a function of N_{agg} and the eccentricity (deviation from spherical shape). The eccentricity (e) is calculated from the principle moments of inertia, according to the following expression:²⁹

$$e = 1 - \frac{I_{\min}}{\langle I \rangle} \quad (7)$$

Here, I_{\min} represents the principal moment of inertia with the smallest magnitude and $\langle I \rangle$ is the average of all three principal moments. Both the corona and core of the micelles are considered in the moment-of-inertia calculations. The eccentricity of a perfectly spherical object is zero, while infinitesimally thin oblate or prolate objects would have eccentricities approaching one.

The micelles observed in the present simulations generally have eccentricities that deviate from zero, which indicates that

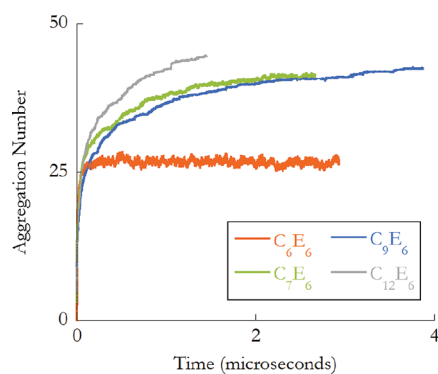


Figure 5. Average micelle aggregation number as a function of time for four surfactant systems, averaged over all replicas. The degree of convergence decreases with increasing hydrophobic tail length, and it appears that more hydrophobic surfactants are approaching larger average aggregation numbers. This qualitative trend is in agreement with the experiment.^{4,84}

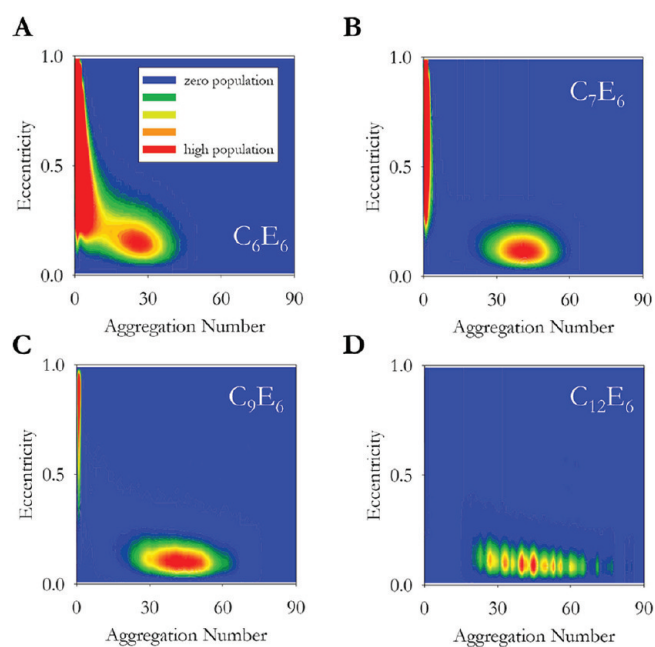


Figure 6. Two-dimensional histograms of micelles as a function of aggregation number and eccentricity (deviation from spherical shape) for four different surfactant systems. In general, aggregates composed of more than 10 monomers are slightly aspherical (eccentricity greater than zero). Trends toward lower eccentricity are observed with increasing hydrocarbon chain length and aggregation number.

they are somewhat aspherical. The larger micelles observed are more spherical, on average, than smaller ones. One would expect this trend to reverse as the micelles transition from spherical to rodlike, but rodlike micelles are not observed in the present simulations.

This trend is also visible in Figure 7, in which one-dimensional slices of the distributions in Figure 6 are shown for clarity. Note that the eccentricity distributions are more sharply peaked near zero for micelles with $N_{\text{agg}} = 55$, compared to those with $N_{\text{agg}} = 30$. In addition, surfactants with longer hydrophobic tails yield more spherical micelles, as can be seen by comparing the one-dimensional eccentricity distributions of C_7E_6 and C_9E_6 at

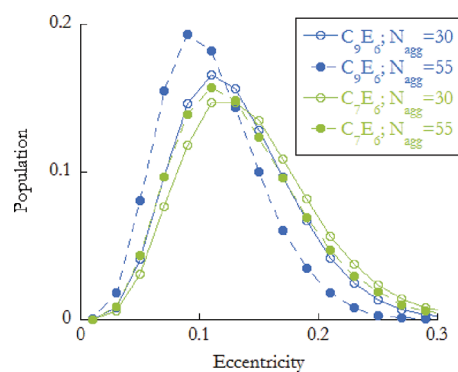


Figure 7. Distribution of eccentricities of micelles differing in aggregation number and the hydrophobic tail length of the component surfactant. As can also be seen in Figure 6, micelles tend to be more spherical (less eccentric) as the length of the hydrophobic chain increases and as the micelles grow larger.

equivalent values of N_{agg} . This trend is in good agreement with the prediction that the radius of a spherical micelle core cannot exceed the length of the hydrocarbon carbon chain.⁴

3.3. Convergence of the Micelle Size Distribution. As noted previously, the peak of the micelle size distribution predicted by our simulations does not change significantly as the hydrophobic tail length increases from 7 to 12, contrary to experimental evidence.^{4,84} The source of this discrepancy can be seen in the time evolution of the mean micelle aggregation number of this series of surfactants, presented in Figure 5. Free monomers are excluded when calculating the mean micelle aggregation number. The more-hydrophilic surfactant shown, C_6E_6 , reaches an average value of 26 in $\sim 0.5 \mu\text{s}$, and subsequently oscillates around this equilibrium size, indicating that this system has likely reached its equilibrium micelle aggregation number distribution. This kinetic equilibrium is supported by the observation of hundreds of micelle fission and fusion processes in these simulations, the kinetics of which will be the subject of a future study. In contrast, after the rapid growth of an initial set of micelles, the average N_{agg} values for the more-hydrophobic systems (C_7E_6 and longer) continue to rise slowly for the duration of the simulations. This indicates that these systems have not reached their equilibrium micelle number distributions, despite the fact that all of the replicas involved in these calculations have a duration of $>1 \mu\text{s}$, and in the C_7E_6 and C_9E_6 simulations, the replica trajectories average $4 \mu\text{s}$ each. It appears that, after their initial formation, micelles grow by slow processes (e.g., micelle fusion), which are not accessible, even by the very efficient GPU-accelerated, CG approach employed in this work. However, note that it is only because of the extreme sampling achieved in this study that we are able to see that the average micelles observed in our simulations of C_7E_6 and C_9E_6 continue to grow for the $4\text{-}\mu\text{s}$ duration of our simulations.

In an effort to assess the degree to which these systems remain out of equilibrium, a system composed of four micelles of 175 C_9E_6 molecules solvated in 96 000 water beads was prepared using the Packmol software package.⁸⁵ This results in the same concentration (127 mM) as the simulations of C_9E_6 discussed in previous sections. Within the first 2 ns of the simulation, each of these large micelles splits into two smaller ones (see Figure 8A). These eight micelles, with an average micelle number of 86.8, remain stable for $2.0 \mu\text{s}$ of MD simulation, despite the fact that no aggregates this large are observed in the original simulations (see

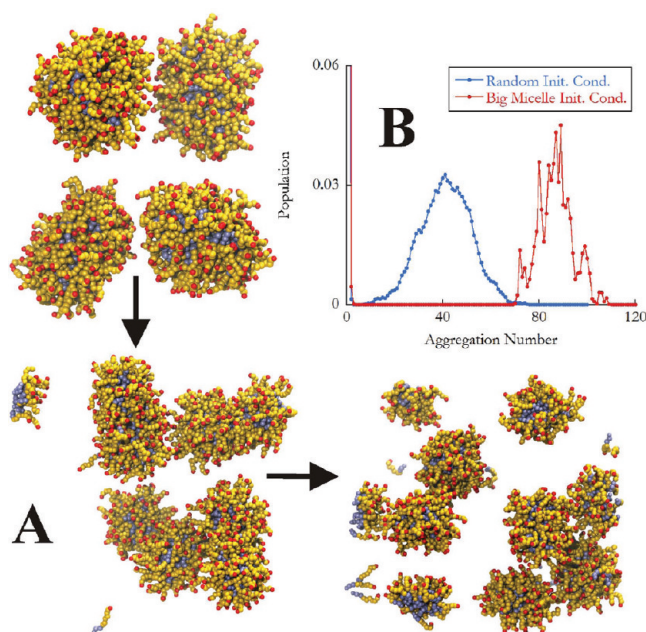


Figure 8. (A) Micelle fission is observed in a simulation that was started from a configuration with four large micelles (175 monomers each) of C_9E_6 (top-left). After these initial fission events (bottom-left), which occurs after only 2 ns of simulations, the resulting micelles remained stable for 2.0 μ s (bottom-right). (B) Aggregation number distribution as determined in simulations started from these “big micelle” initial conditions (red), compared with the distribution determined from random, unaggregated initial conditions (blue) at the same total concentration of monomer.

Figure 8B). The discrepancy between the size of stable micelles observed in simulations starting from different initial conditions indicates that advanced sampling techniques, such as replica exchange, are necessary to overcome the large barriers associated with micelle fission and fusion. Such an approach has previously been used in conjunction with GPU-accelerated coarse grained MD in a study of protein–protein interactions.⁷⁴ In future work, we will incorporate advanced sampling techniques into our GPU-accelerated coarse-grained MD scheme to more fully sample the equilibrium distribution of micellar solutions.

3.4. Determination of Critical Micelle Concentration. Critical micelle concentrations (CMCs), which are the concentrations above which particular surfactants will form micelles, were predicted from our simulations of micellar solutions at ~ 128 mM. This is done by measuring the free monomer concentration in these micellar solutions, which analytic theory predicts to be equal to the CMC.³ The theoretically predicted CMCs of the series of C_xE_6 surfactants are given in Figure 9, along with experimentally determined CMCs for comparison.^{84,86,87} As is known, there is tremendous variability in experimental measurements of the CMC, and the experimental measurements of the CMC reported here vary by a factor of 2 in some cases. The strong decrease in the CMC with increasing hydrophobic tail length predicted by the simulations is in good qualitative agreement with experimental results, and the values of the CMC predicted for the more-hydrophilic surfactant systems (C_6E_6 and C_7E_6) are in good quantitative agreement with the experiment. However, for longer hydrophobic tails, the CMC predicted by MD simulation is systematically too small. The simulated CMC deviates from the experimental CMC by a factor of 10 in the worst case ($C_{12}E_6$). It is possible that more-accurate

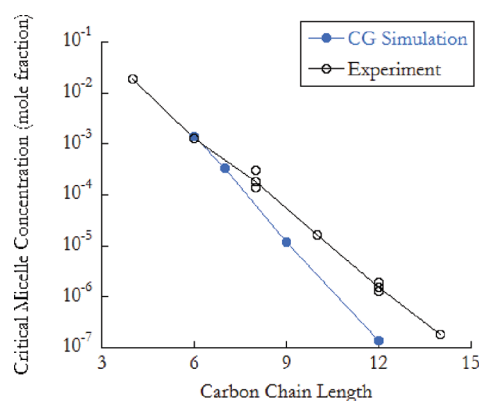


Figure 9. Calculated critical micelle concentrations (CMC) as a function of hydrocarbon chain length are shown in blue, with experimental data^{84,86,87} shown in black for comparison. Simulated CMCs are determined from the free monomer concentration of micellar solutions, as described in the text. Agreement with experiment is excellent for short hydrocarbon chains (C_6E_6 and C_7E_6), and the qualitative trend of decreasing CMC with increasing hydrophobicity is reproduced.

results would be obtained in these cases if the simulations were conducted at concentrations closer to the CMC; however, unfortunately, such simulations are not feasible, because of the sheer volume of solvent that would be required to simulate even a single micelle at such low monomer concentration.

Unlike the micelle size distributions previously discussed, the free monomer concentrations used to predict the CMCs appear to have reached converged values relatively early in the simulations. As can be seen in Figure 10, the average number of free monomers observed in each frame of MD trajectory data converges within the first several hundred nanoseconds of simulation. Indeed, the free monomer concentrations reported here appear to be converged.

The dependence of the CMC on the PEG chain length was also calculated. The aggregation number distributions for a series of three surfactant systems of the form C_6E_y are shown in Figure 11. For $y = 8$, only relatively small aggregates are observed, and no minimum is observed in the micelle size distribution, so it is difficult to discriminate between micelles and free monomers. However, there is a strong dependence of the average size of micelles on the PEG chain length, so it is easy to discriminate between micelles and free monomers for C_6E_6 and C_6E_4 . In comparison to the length of the hydrophobic tail, which has a very strong effect on the CMC, the length of the hydrophilic PEG chain correlates relatively weakly with the CMC. This trend can be seen in the inset of Figure 11 and is in good agreement with experimental studies of similar PEG surfactant systems, which show a slight increase in the CMC with increasing PEG chain length in surfactants of the forms C_8E_y and $C_{10}E_y$,⁸⁷ as well as other simulations.²¹

3.5. Behavior Near the Critical Micelle Concentration. Analytic theory³ and phenomenological models⁴ predict the behavior of surfactant solutions at concentrations surrounding the CMC. At concentrations above the CMC, the concentration of free monomers is expected to be approximately the CMC and remain constant with increasing total concentration. For lower concentration solutions, it is predicted that all monomers will exist as free monomers. However, Vold has suggested that it is more correct to think of the surfactants in solutions approaching the CMC not as free monomers, but as small pre-micellar aggregates.⁸⁰ Many experiments support this suggestion,^{88–94}

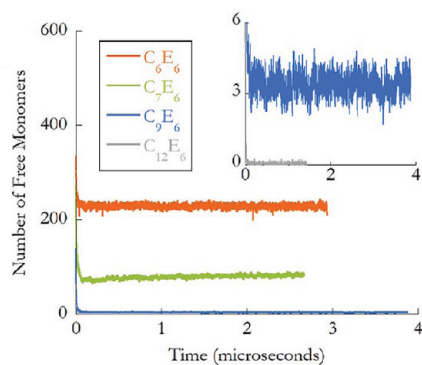


Figure 10. Average number of instantaneous free monomers from simulations of five surfactant systems at the same concentration as a function of time. Values are averaged over all replicas. The inset shows two surfactant systems with a smaller range of numbers of free monomers for clarity. The number of free monomers converges to an average value within $<0.2 \mu\text{s}$ in all cases.

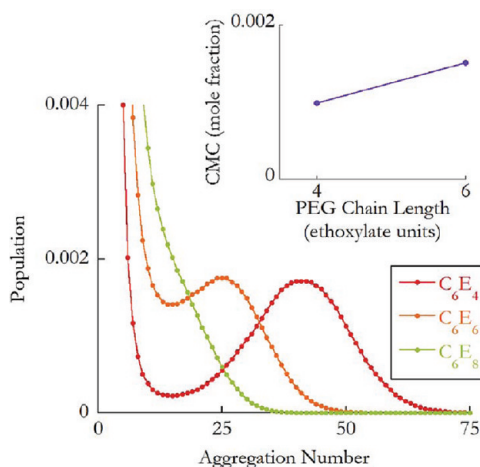


Figure 11. Aggregation number distribution for a series of surfactants differing in the length of the hydrophilic PEG chain, but with constant hydrophobic tail length and concentration. While the PEG chain length has a significant effect on the peak aggregation number, the fraction of monomers forming micelles is relatively constant, indicating that the CMC is relatively insensitive to PEG chain length. No CMC can be determined for C_6E_8 , because no peak is observed in the aggregation number (though a shoulder is observed for N_{agg} in the range of 15–25, which corresponds to small, micelle-like aggregates). The CMCs determined for the other two surfactant systems are plotted in the inset. The CMC increases slightly with increasing PEG chain length. This qualitative trend is in good agreement with experimental data.⁸⁷

including a recent dynamic light scattering experiment, demonstrating the existence of such aggregates in PEG surfactant solutions.⁸¹ In a previous publication,⁹⁵ the behavior of surfactant solutions at concentrations surrounding the CMC and the prevalence of pre-micellar aggregates was investigated by conducting sets of simulations spanning a range of concentrations surrounding the CMC for two different PEG surfactants, C_7E_6 and C_4E_6 . It was reported that, in C_7E_6 solutions above the CMC, the free monomer concentration (which, in our analysis, is assumed to be a reasonable predictor of the CMC) does not remain constant as the total concentration increases. In fact, it decreases with total concentration, as can be seen in Figure 12. This slow decrease in the free monomer concentration

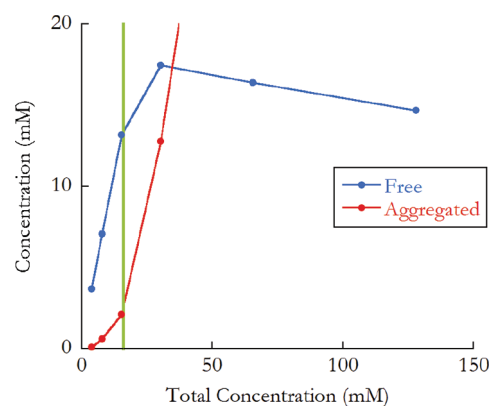


Figure 12. Concentration of unaggregated monomers (those which compose aggregates with $N_{\text{agg}} = 1$; blue) and aggregated monomers (those which compose aggregates with $N_{\text{agg}} > 1$; red) as a function of total C_7E_6 concentration in the neighborhood of the CMC. The CMC, as determined for this model—16.1 mM—is marked by a vertical green line. Two notable deviations from ideal behavior are observed.⁹⁵ Above the CMC, the free monomer concentration plateaus, as predicted by analytic theory.⁴ However, there is a slight, but unexpected decrease as the total concentration increases, which suggests that a more-accurate CMC can be predicted by calculating the free monomer concentration at a concentration as near to the CMC as possible. In addition, aggregates are observed at a concentration below the CMC, with 8% of monomers aggregated at a concentration of approximately half the CMC.

with increasing total concentration supports the suggestion that a more-accurate CMC can be predicted by simulating at a concentration nearer to the CMC than is feasible for the more-hydrophobic C_9E_6 and C_{12}E_6 systems, where the largest deviations of our calculated CMCs from experimental results were observed.

Another noteworthy deviation from ideal behavior, the existence of pre-micelles, was observed in C_7E_6 solutions below the CMC.⁹⁵ In simulations at a concentration less than half of the CMC, 8% of surfactant molecules aggregated, as can be seen in Figure 12. Dimeric structures were observed to be similarly stable to micelles, and significant populations of pre-micelles composed of between three and nine monomers also were observed.

In Figure 13, we illustrate typical morphologies of C_7E_6 pre-micelles from a 128 mM solution, with variations in both aggregation number and spherical eccentricity. Along the columns are representative depictions of pre-micelles ranging in shape from less eccentric ($e = 0.125$) to highly eccentric ($e = 0.65$), and along the rows, we show pre-micelles with aggregation numbers ranging trimers, to pentamers and nonamers, at a constant eccentricity. The diagonal extending from $\{e = 0.65, N_{\text{agg}} = 3\}$ to $\{e = 0.125, N_{\text{agg}} = 9\}$ roughly corresponds to the most probable eccentricity at a given aggregate size. For all equivalently sized systems, increasing the eccentricity generally results in elongation of the pre-micelle along its major principal axis, while retaining symmetry along the other two. It is noteworthy that we observed a lack of spherical pre-micellar structures, as typically used, as a matter of convenience, in the formulation of theoretical models for predicting pre-micellar thermodynamics.⁹⁶

Deviations from ideal behavior are more severe in aqueous C_4E_6 , as can be seen in Figure 14, which presents aggregation number distributions calculated at several concentrations in the vicinity of the experimentally reported CMC, 760 mM.⁸⁶ As

the total concentration increases, no peak in the aggregation number distribution develops, despite the fact that aggregates of greater than 100 monomers are observed as the total surfactant concentration approaches 1 M. Similar nonideal behavior is observed if one considers the relationship between the concentrations of smaller aggregates and free monomers and the total surfactant concentration; no sudden leveling of the concentration of free monomers (inset of Figure 14B) or premicellar aggregates (Figure 14B) at a critical concentration is observed. Instead, the concentration of free monomers smoothly transitions from increasing with total surfactant concentration to decreasing at ~ 500 mM total concentration, in contrast to the sudden change in behavior predicted at the CMC and observed in our simulations of C_7E_6 . This indicates that C_4E_6 is not strongly micelle forming and suggests that no CMC can be defined for C_4E_6 in water. Although a CMC for C_4E_6 has previously been inferred from surface tension experiments,⁸⁶

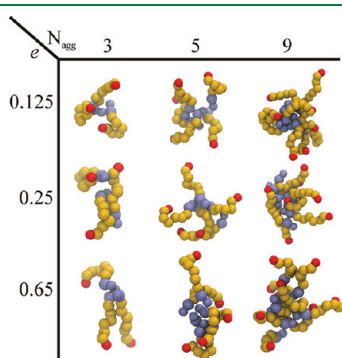


Figure 13. Morphology of premicellar aggregates as a function of aggregation number ($N_{\text{agg}} = 3, 5, 9$) and spherical eccentricity ($e = 0.125, 0.25, 0.65$; see eq 7). The diagonal extending from the lower left to the top right follows the peak in the respective system's eccentricity distribution. By increasing the eccentricity (columns), premicellar aggregates tend to elongate along a major principal axis for all three system. Note that perfectly spherical premicelles are not observed at any aggregation size.

no raw data was reported, so it is impossible to assess the degree to which the change in behavior at the reported CMC is sudden. Given the large aggregates observed in our simulations, it would not be surprising if the surface tension of C_4E_6 does not depend strongly on the concentration around the experimentally determined CMC (760 mM), even if no critical behavior is observable. To the authors' knowledge, no other measurement of the CMC of C_4E_6 has been reported in the literature.

4. CONCLUSIONS

In this paper, we presented a study using GPU-accelerated coarse-grained (CG) molecular dynamics (MD) simulations to investigate the properties of micellar solutions of nonionic surfactants. In order to carry out this study, a GPU-accelerated implementation of the SDK force field¹ was added to the HOOMD-Blue GPU-accelerated MD software package.² The performance of a 25 000 particle MD simulation on a single GPU was comparable to that obtained using the LAMMPS MD software package^{82,83} running on an optimally sized computer cluster comprising tens of CPU cores. All code developed and used in this work is available for free download with the current production release of the HOOMD-Blue GPU-accelerated MD software package.⁷⁹

The combined efficiency afforded us by the CG model and GPU hardware allowed us to undertake an unprecedented study of surfactant self-assembly, collecting a total of 0.6 ms of trajectory data in 200 replica simulations. From this data, we calculated and reported many different thermodynamic and structural properties. Converged critical micelle concentrations (CMCs) exhibited the qualitatively correct dependence on the length of the hydrophobic tail and hydrophilic PEG headgroup. Quantitative agreement with experiment was achieved for more-hydrophilic surfactant systems (C_6E_6 and C_7E_6), while the CMCs of more hydrophobic surfactants were underestimated significantly. For C_7E_6 , it has been previously reported that the free monomer concentration decreases as the total monomer concentration increases beyond the CMC,⁹⁵ in contrast to widely used models that predict it to remain relatively constant. This trend suggests that errors in the CMCs of the more hydrophobic

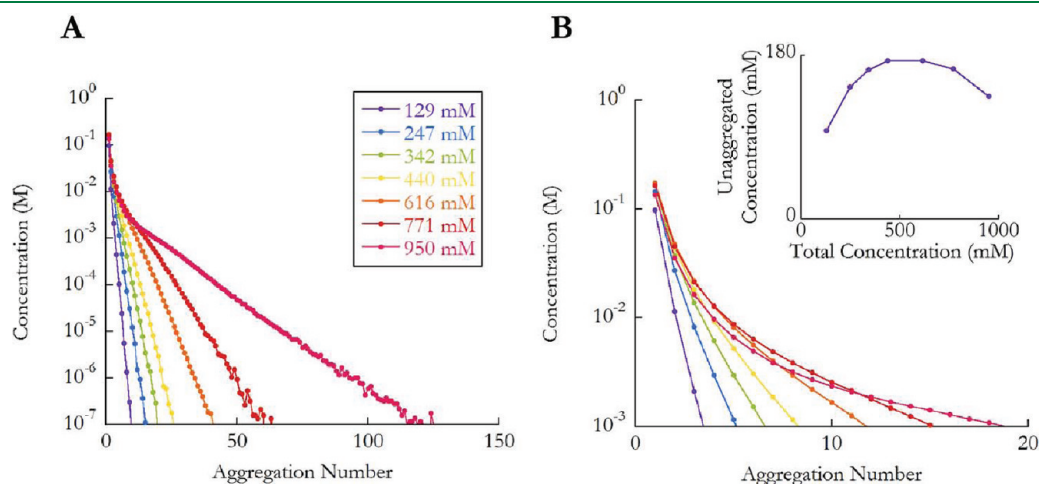


Figure 14. (A) Distribution of micelle aggregation numbers for aqueous C_4E_6 surfactant systems at various total surfactant concentrations. (B) Enlarged view showing the concentration of small premicellar aggregates. The concentration of unaggregated monomers (those which compose aggregates with $N_{\text{agg}} = 1$), as a function of total monomer concentration in the neighborhood of a previously reported CMC (760 mM),³⁰ is shown in the inset of Panel B. Unlike C_7E_6 (see Figure 12), and despite the fact that large aggregates exist at high concentration, C_4E_6 does not exhibit a sharp change in behavior at a particular concentration (CMC). In addition, its aggregation number distribution develops a peak at high concentration more similar to the more-hydrophobic surfactants (see Figure 4).

surfactants may result from the fact that the simulations are carried out at concentrations much above the CMC. For the most-hydrophilic surfactant studied, C_4E_6 , no sharp change in behavior is seen at any concentration, suggesting that no CMC can be defined for this system.

The average micelle aggregation numbers determined from our simulations increase with hydrophobic chain length, which is in agreement with experimental results. However, even with the extreme sampling achieved in this work, equilibrium micelle size distributions are achieved for only the most-hydrophilic surfactant systems. In fact, it is only because of the large trajectory data set collected that the slow micellar growth that continues for the duration of our simulations can be observed at all. In future work, we will examine the dynamics of this growth and employ advanced sampling techniques in conjunction with our GPU-accelerated CG scheme to attempt to sample true equilibrium distributions for the more challenging hydrophobic surfactant systems.

■ ASSOCIATED CONTENT

S Supporting Information. An additional Supporting Information file is available, containing the following: tables of force-field parameters; radius of gyration distributions for various micellar systems; plots showing the sensitivity of the micelle size distribution, with respect to the clustering cutoff; and a sample HOOMD input file for running a PEG surfactant system with the SDK model. This information is available free of charge via the Internet at <http://pubs.acs.org>.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: mlklein@temple.edu.

Author Contributions

[†]D.N.L. and B.G.L. contributed equally to this work.

■ ACKNOWLEDGMENT

The authors gratefully acknowledge many useful discussions with Arben Jusufi, Joshua Anderson, Alex Travasset, Athanasios Panagiotopoulos, Samantha Sanders, and Steve Barr. We are also extremely grateful to Joshua Anderson and Paul Navratil for technical assistance. This work was supported by the National Science Foundation (under Grant Nos. CHE-09-46358 and CNS-09-58854) and through TeraGrid resources provided by the Texas Advanced Computing Center, National Center for Supercomputer Applications, and National Institute for Computational Science (under Grant Nos. TG-MCA93S020 and TG-ASC090088). R.D. and M.L.K. would also like to thank The Procter & Gamble company for generous support.

■ REFERENCES

- (1) Shinoda, W.; DeVane, R.; Klein, M. L. *Mol. Simul.* **2007**, *33*, 27–36.
- (2) Anderson, J. A.; Lorenz, C. D.; Travasset, A. *J. Comput. Phys.* **2008**, *227*, 5342–5359.
- (3) Israelachvili, J. N. *Intermolecular and Surface Forces*; Harcourt Brace: London, 1992.
- (4) Tanford, C. *The Hydrophobic Effect: Formation of Micelles and Biological Membranes*; John Wiley & Sons: New York, 1973.
- (5) Xia, Y. N.; Whitesides, G. M. *Annu. Rev. Mater. Sci.* **1998**, *28*, 153–184.

- (6) Kato, T.; Mizoshita, N.; Kishimoto, K. *Angew. Chem., Int. Ed.* **2006**, *45*, 38–68.
- (7) Klein, M. L.; Shinoda, W. *Science* **2008**, *321*, 798–800.
- (8) Marrink, S. J.; de Vries, A. H.; Tieleman, D. P. *Biochim. Biophys. Acta* **2009**, *1788*, 149–168.
- (9) Watanabe, K.; Ferrario, M.; Klein, M. L. *J. Phys. Chem.* **1988**, *92*, 819–821.
- (10) Watanabe, K.; Klein, M. L. *J. Phys. Chem.* **1989**, *93*, 6897–6901.
- (11) Smit, B.; Hilbers, P. A. J.; Esselink, K.; Rupert, L. A. M.; van Os, N. M.; Schlijper, A. G. *Nature* **1990**, *348*, 624–625.
- (12) Smit, B.; Hilbers, P. A. J.; Esselink, K.; Rupert, L. A. M.; van Os, N. M.; Schlijper, A. G. *J. Phys. Chem.* **1991**, *95*, 6361–6368.
- (13) Karaborni, S.; Esselink, K.; Hilbers, P. A. J.; Smit, B.; Karthaus, J.; van Os, N. M.; Zana, R. *Science* **1994**, *266*, 254–256.
- (14) Karaborni, S.; Smit, B. *Curr. Opin. Colloid Interface Sci.* **1996**, *1*, 411–415.
- (15) Abel, S.; Dupradeau, F.-Y.; Raman, E. P.; MacKerell, A. D.; Marchi, M. J. *J. Phys. Chem.* **2011**, *115B*, 487–499.
- (16) Amani, A.; York, P.; de Waard, H.; Anwar, J. *Soft Matter* **2011**, *7*, 2900–2908.
- (17) Bruce, C. D.; Berkowitz, M. L.; Perera, L.; Forbes, M. D. E. *J. Phys. Chem.* **2002**, *106B*, 3788–3793.
- (18) Fujiwara, S.; Itoh, T.; Hashimoto, M.; Horiuchi, R. *J. Chem. Phys.* **2009**, *130*, 144901.
- (19) Gao, J.; Ge, W.; Hu, G.; Li, J. *Langmuir* **2005**, *21*, 5223–5229.
- (20) Garde, S.; Yang, L.; Dordick, J. S.; Paulaitis, M. E. *Mol. Phys.* **2002**, *100*, 2299–2306.
- (21) Jusufi, A.; Sanders, S.; Klein, M. L.; Panagiotopoulos, A. Z. *J. Phys. Chem.* **2011**, *115B*, 990–1001.
- (22) Lazaridis, T.; Mallik, B.; Chen, Y. *J. Phys. Chem.* **2005**, *109B*, 15098–15106.
- (23) Marrink, S. J.; Tieleman, D. P.; Mark, A. E. *J. Phys. Chem.* **2000**, *104B*, 12165–12173.
- (24) Sammalkorpi, M.; Sanders, S.; Panagiotopoulos, A. Z.; Karttunen, M.; Haataja, M. *J. Phys. Chem.* **2011**, *115B*, 1403–1410.
- (25) Stephenson, B. C.; Stafford, K. A.; Beers, K. J.; Blankschtein, D. *J. Phys. Chem.* **2008**, *112B*, 1641–1656.
- (26) Tieleman, D. P.; van der Spoel, D.; Berendsen, H. J. C. *J. Phys. Chem.* **2000**, *104B*, 6380–6388.
- (27) Chowdhary, J.; Ladanyi, B. M. *J. Phys. Chem.* **2009**, *113B*, 15029–15039.
- (28) Salaniwal, S.; Cui, S. T.; Cochran, H. D.; Cummings, P. T. *Langmuir* **2001**, *17*, 1773–1783.
- (29) Senapati, S.; Keiper, J. S.; DeSimone, J. M.; Wignall, G. D.; Melnichenko, Y. B.; Frielinghaus, H.; Berkowitz, M. L. *Langmuir* **2002**, *18*, 7371–7376.
- (30) Denham, N.; Holmes, M. C.; Zvelindovsky, A. V. *J. Phys. Chem.* **2011**, *115B*, 1385–1393.
- (31) Drouffe, J. M.; Maggs, A. C.; Leibler, S. *Science* **1991**, *254*, 1352–1356.
- (32) Goetz, R.; Lipowsky, R. *J. Chem. Phys.* **1998**, *108*, 7397–7409.
- (33) He, X.; Shinoda, W.; DeVane, R.; Anderson, K. L.; Klein, M. L. *J. Chem. Phys. Lett.* **2010**, *487*, 71–76.
- (34) Izvekov, S.; Voth, G. A. *J. Phys. Chem.* **2009**, *113B*, 4443–4455.
- (35) Shinoda, W.; DeVane, R.; Klein, M. L. *Soft Matter* **2008**, *4*, 2454–2462.
- (36) Shinoda, W.; DeVane, R.; Klein, M. L. *J. Phys. Chem.* **2010**, *20*, 6836–6849.
- (37) Baoukina, S.; Monticelli, L.; Risselada, H. J.; Marrink, S. J.; Tieleman, D. P. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 10803–10808.
- (38) Sammalkorpi, M.; Karttunen, M.; Haataja, M. *J. Am. Chem. Soc.* **2008**, *130*, 17977–17980.
- (39) Anderson, J. A.; Travasset, A. *Macromolecules* **2006**, *39*, 5143–5151.
- (40) Phillips, C. L.; Iacovella, C. R.; Glotzer, S. C. *Soft Matter* **2010**, *6*, 1693–1703.
- (41) Bowers, K. J.; Chow, E.; Xu, H.; Dror, R. O.; Eastwood, M. P.; Gregersen, B. A.; Klepeis, J. L.; Kolossvary, I.; Moraes, M. A.; Sacerdoti,

- F. D.; Salmon, J. K.; Shan, Y.; Shaw, D. E. In *Proceedings of the 2006 ACM/IEEE Conference on Supercomputing*; IEEE Press: Piscataway, NJ, 2006; pp 11–17.
- (42) Darden, T.; York, D.; Pedersen, L. J. *Chem. Phys.* **1993**, *98*, 10089–10092.
- (43) Tuckerman, M.; Berne, B. J.; Martyna, G. J. *J. Chem. Phys.* **2008**, *97*, 1990.
- (44) Christen, M.; Van Gunsteren, W. F. *J. Comput. Chem.* **2008**, *29*, 157–166.
- (45) Liwo, A.; Czaplowski, C.; Oldziej, S.; Scheraga, H. A. *Curr. Opin. Struct. Biol.* **2008**, *18*, 134–139.
- (46) Shaw, D. E. *Commun. ACM* **2008**, *51*, 91–97.
- (47) Izvekov, S.; Voth, G. A. *J. Phys. Chem.* **2005**, *109B*, 2469–2473.
- (48) Marrink, S. J.; Risselada, H. J.; Yefimov, S.; Tieleman, D. P.; de Vries, A. H. *J. Phys. Chem.* **2007**, *111B*, 7812–7824.
- (49) Owens, J. D.; Houston, M.; Luebke, D.; Green, S.; Stone, J. E.; Phillips, J. C. *Proc. IEEE* **2008**, *96*, 879–899.
- (50) Anderson, A. G.; Goddard, I. W. A.; Schroder, P. *Comput. Phys. Commun.* **2007**, *177*, 298–306.
- (51) Asadchev, A.; Allada, V.; Felder, J.; Bode, B. M.; Gordon, M. S.; Windus, T. L. *J. Chem. Theory Comput.* **2010**, *6*, 696–704.
- (52) Bauer, B. A.; Davis, J. E.; Tauber, M.; Patel, S. J. *Comput. Chem.* **2011**, *32*, 375–385.
- (53) Brown, P.; Woods, C.; McIntosh-Smith, S.; Manby, F. R. *J. Chem. Theory Comput.* **2008**, *4*, 1620–1626.
- (54) Brown, M.; Wang, P.; Plimpton, S. J.; Tharrington, A. N. *Comput. Phys. Commun.* **2011**, *182*, 898–911.
- (55) Eastman, P.; Pande, V. S. *J. Chem. Theory Comput.* **2010**, *6*, 434–437.
- (56) Elsen, E.; Houston, M.; Vishal, V.; Darve, E.; Hanrahan, P.; Pande, V. In *Proceedings of the 2006 ACM/IEEE Conference on Supercomputing*; IEEE Press: Piscataway, NJ, 2006; p 188.
- (57) Friedrichs, M. S.; Eastman, P.; Vaidyanathan, V.; Houston, M.; LeGrand, S.; Beberg, A. L.; Ensign, D. L.; Brums, C. M.; Pande, V. S. *J. Comput. Chem.* **2009**, *30*, 864–872.
- (58) Harvey, M. J.; Giupponi, G.; De Fabritiis, G. *J. Chem. Theory Comput.* **2009**, *5*, 1632–1639.
- (59) Harvey, M. J.; De Fabritiis, G. *J. Chem. Theory Comput.* **2009**, *5*, 2371–2377.
- (60) Jha, P. K.; Sknepnek, R.; Guerrero-Garcia, G. I.; de la Cruz, M. O. *J. Chem. Theory Comput.* **2010**, *6*, 3058–3065.
- (61) Narumi, T.; Yasuoka, K.; Taiji, M.; Hoefinger, S. *J. Comput. Chem.* **2009**, *30*, 2351–2357.
- (62) Olivares-Amaya, R.; Watson, M. A.; Edgar, R. G.; Vogt, L.; Shao, Y.; Aspuru-Guzik, A. *J. Chem. Theory Comput.* **2010**, *6*, 135–144.
- (63) Peng, L.; Nomura, K.; Oyakawa, T.; Kalia, R. K.; Nakano, A.; Vashishta, P. *14th International Euro-Par Conference*; Springer–Verlag: Berlin, Germany, 2008; pp 763–777.
- (64) Phillips, J. C.; Stone, J. E.; Schulten, K. In *Proceedings of the 2008 ACM/IEEE Conference on Supercomputing*; IEEE Press: Piscataway, NJ, 2008; p 1–9.
- (65) Rodrigues, C. I.; Hardy, D. J.; Stone, J. E.; Schulten, K.; Hwu, W.-M. W. In *Proceedings of the 2008 Conference on Computing Frontiers*; ACM: New York, 2008; pp 273–282.
- (66) Stone, J. E.; Saam, J.; Hardy, D. J.; Vandivort, K. L.; Hwu, W.-M. W.; Schulten, K. In *Proceedings of the 2nd Workshop on General Purpose Processing on Graphics Processing Units*; ACM International Conference Proceeding Series; ACM: Washington, DC, 2009; Vol. 383; pp 9–18.
- (67) Stone, J. E.; Phillips, J. C.; Freddolino, P. L.; Hardy, D. J.; Trabuco, L. G.; Schulten, K. *J. Comput. Chem.* **2007**, *28*, 2618–2640.
- (68) Sukhwani, B.; Herbordt, M. C. *Proceedings of the 2nd Workshop on General Purpose Processing on Graphics Processing Units*; ACM: Washington, DC, 2009; Vol. 383; pp 19–27.
- (69) Ufimtsev, I. S.; Martinez, T. J. *Comput. Sci. Eng.* **2008**, *10*, 26–34.
- (70) Ufimtsev, I. S.; Martinez, T. J. *J. Chem. Theory Comput.* **2009**, *5*, 1004–1015.
- (71) Vogt, L.; Olivares-Amaya, R.; Kermes, S.; Shao, Y.; Amador-Bedolla, C.; Aspuru-Guzik, A. *J. Phys. Chem.* **2008**, *112A*, 2049–2057.
- (72) Watson, M. A.; Olivares-Amaya, R.; Edgar, R. G.; Aspuru-Guzik, A. *Comput. Sci. Eng.* **2010**, *12*, 40–51.
- (73) Yasuda, K. *J. Chem. Theory Comput.* **2008**, *4*, 1230–1236.
- (74) Tunbridge, I.; Best, R. B.; Gain, J.; Kuttel, M. M. *J. Chem. Theory Comput.* **2010**, *6*, 3588–3600.
- (75) *Texas Advanced Computing Center: Visualization*, <http://www.tacc.utexas.edu/resources/visualization/.07/01/2011>.
- (76) *Intel 64 T Linux Cluster Lincoln*, <http://www.ncsa.illinois.edu/UserInfo/Resources/Hardware/Intel64TeslaCl>.
- (77) *TOP500 Supercomputing Sites*, <http://www.top500.org/lists/2010/11.12/01/2010>, accessed November 2010.
- (78) He, X.; Shinoda, W.; DeVane, R.; Klein, M. L. *Mol. Phys.* **2010**, *108*, 2007–2020.
- (79) *HOOMD-Blue Download*, <http://codeblue.umich.edu/hoomd-blue/download.html> (accessed July 1, 2011).
- (80) Vold, M. *Langmuir* **1992**, *8*, 1082–1085.
- (81) Lee, Y.-C.; Liu, H.-S.; Lin, S.-Y.; Huang, H.-F.; Wang, Y.-Y.; Chou, L.-W. *J. Chin. Inst. Chem. Eng.* **2008**, *39*, 75–83.
- (82) Plimpton, S. J. *Comput. Phys.* **1995**, *117*, 1–19.
- (83) *LAMMPS*, <http://lammps.sandia.gov>, accessed July 1, 2011.
- (84) Balmbra, R. R.; Clunie, J. S.; Corkill, J. M.; Goodman, J. F. *Trans. Faraday Soc.* **1964**, *60*, 979–985.
- (85) Martinez, L.; Andrade, R.; Birgin, E. G.; Martinez, J. M. *J. Comput. Chem.* **2009**, *30*, 2157–2164.
- (86) Elworthy, P. H.; Florence, A. T. *Kolloid Z. Z. Polym.* **1964**, *195*, 23–27.
- (87) Corkill, J. M.; Goodman, J. F.; Harrol, S. P. *Trans. Faraday Soc.* **1964**, *60*, 202–207.
- (88) Niu, S.; Gopidas, K. R.; Turro, N. J.; Gabor, G. *Langmuir* **1992**, *8*, 1271–1277.
- (89) Menger, F. M.; Littau, C. A. *J. Am. Chem. Soc.* **1993**, *115*, 10083–10090.
- (90) Sakai, T.; Kaneko, Y.; Tsujii, K. *Langmuir* **2006**, *22*, 2039–2044.
- (91) Cui, X.; Mao, S.; Liu, M.; Yuan, H.; Du, Y. *Langmuir* **2008**, *24*, 10771–10775.
- (92) Barnadas-Rodriguez, R.; Estelrich, J. *J. Phys. Chem.* **2009**, *113B*, 1972–1982.
- (93) Sowmiya, M.; Tiwari, A. K.; Saha, S. K. *J. Colloid Interface Sci.* **2010**, *344*, 97–104.
- (94) Beija, M.; Fedorov, A.; Charreyre, M. T.; Marinho, J. M. G.; Jose, M. G. *J. Phys. Chem.* **2010**, *114B*, 9977–9986.
- (95) LeBard, D. N.; Levine, B. G.; DeVane, R.; Shinoda, W.; Klein, M. L. *Chem. Phys. Lett.* **2011**, submitted for publication. (See Supporting Information.)
- (96) Hadjiivanova, R.; Diamant, H. *J. Phys. Chem. B* **2007**, *111B*, 8854–8859.

Methods for Computing Accurate Atomic Spin Moments for Collinear and Noncollinear Magnetism in Periodic and Nonperiodic Materials

Thomas A. Manz* and David S. Sholl*

School of Chemical and Biomolecular Engineering, Georgia Institute of Technology, 311 Ferst Drive N.W., Atlanta, Georgia 30332-0100, United States

S Supporting Information

ABSTRACT: The partitioning of electron spin density among atoms in a material gives atomic spin moments (ASMs), which are important for understanding magnetic properties. We compare ASMs computed using different population analysis methods and introduce a method for computing density derived electrostatic and chemical (DDEC) ASMs. Bader and DDEC ASMs can be computed for periodic and nonperiodic materials with either collinear or noncollinear magnetism, while natural population analysis (NPA) ASMs can be computed for nonperiodic materials with collinear magnetism. Our results show Bader, DDEC, and (where applicable) NPA methods give similar ASMs, but different net atomic charges. Because they are optimized to reproduce both the magnetic field and the chemical states of atoms in a material, DDEC ASMs are especially suitable for constructing interaction potentials for atomistic simulations. We describe the computation of accurate ASMs for (a) a variety of systems using collinear and noncollinear spin DFT, (b) highly correlated materials (e.g., magnetite) using DFT+U, and (c) various spin states of ozone using coupled cluster expansions. The computed ASMs are in good agreement with available experimental results for a variety of periodic and nonperiodic materials. Examples considered include the antiferromagnetic metal organic framework $\text{Cu}_3(\text{BTC})_2$, several ozone spin states, mono- and binuclear transition metal complexes, ferri- and ferro-magnetic solids (e.g., Fe_3O_4 , Fe_3Si), and simple molecular systems. We briefly discuss the theory of exchange-correlation functionals for studying noncollinear magnetism. A method for finding the ground state of systems with highly noncollinear magnetism is introduced. We use these methods to study the spin-orbit coupling potential energy surface of the single molecule magnet $\text{Fe}_4\text{C}_{40}\text{H}_{52}\text{N}_4\text{O}_{12}$, which has highly noncollinear magnetism, and find that it contains unusual features that give a new interpretation to experimental data.

1. INTRODUCTION

Common sources of magnetism in materials include unpaired electrons, nuclear spins, orbital moments, and electric currents. Here, we focus on the magnetic field component $\vec{B}^{\text{spin}}(\vec{r})$ due to unpaired electrons. In chemical systems containing unpaired electrons, the assignment of an atomic spin moment (ASM) to each atom serves several important purposes. First, this helps quantify the type of magnetic state (ferromagnetic, antiferromagnetic, ferrimagnetic, etc.) of a material. By allowing one to theoretically study atomistic changes in magnetism as a function of chemical composition, ASMs facilitate the understanding of existing materials and the design of new materials. ASMs computed from quantum chemistry calculations can be used to develop interaction potentials for atomistic simulations of magnetism that facilitate the study of larger systems and statistical ensembles. For example, atomistic simulations using Ising and other lattice-based models have been used to study magnetic phase transitions, magnetic domain boundaries, spin spirals, and other complex phenomena.^{1–3} To serve all of these purposes, ASMs should be simultaneously optimized to reproduce $\vec{B}^{\text{spin}}(\vec{r})$ and the chemical states of atoms in a material. Here, we describe how ASMs optimized for these uses can be computed from the spin density of a quantum chemistry calculation. We compare these results to both experimental data and other methods for computing ASMs.

Common methods for computing ASMs include Hirshfeld, Mulliken, natural, and Bader population analysis.^{4–10} Integrating

the total spin density within an effective atomic sphere radius is another common method for computing ASMs, but this has the disadvantage that regions between atomic spheres are not assigned to any atom, so the ASMs do not sum to the total spin moment of the unit cell.¹¹ The Hirshfeld method is known to give net atomic charges (NACs) too small in magnitude.^{12–14} Basis set sensitivity is a key disadvantage of Mulliken population analysis.^{7,8} Because of these well-documented problems, we do not consider Hirshfeld and Mulliken analysis further in this article. Natural population analysis (NPA) addresses the basis set issue,^{7,8} but it is not readily available for periodic systems or noncollinear magnetism. In some cases, Mulliken populations can also be stabilized by projection onto a minimal basis set (MBS).^{8,15} Bader analysis partitions space into nonoverlapping atomic volumes and assigns all of the spin density within an atomic volume to that atom.^{4,5,9} Bader analysis does not directly depend on the basis set and can be readily applied to periodic and nonperiodic systems with collinear or noncollinear magnetism. Although Bader NACs are chemically meaningful, they do not accurately reproduce the electrostatic potential, $V(\vec{r})$, outside the electron distribution, which limits their use for force-field development.¹⁴ As shown in the Results below, the NPA and Bader methods give reliable ASMs.

Received: August 2, 2011

Published: September 23, 2011

Previously, we introduced density derived electrostatic and chemical (DDEC) charges that are simultaneously optimized to reproduce both the chemical states of atoms and $V(\vec{r})$ outside a material's electron distribution, which makes them suitable for use in force fields and for studying electron transfer during chemical reactions.¹⁴ DDEC NACs have been used to model gas adsorption in metal organic frameworks and to quantify the amount of charge transfer in molecular zippers.^{16,17} Here, we extend the DDEC method to the computation of ASMs in both collinear and noncollinear magnetism. The DDEC NACs and ASMs are self-consistently derived from the same atomic charge density distributions. These ASMs are simultaneously optimized to reproduce $\vec{B}^{\text{spin}}(\vec{r})$ and the chemical states of atoms in a material, which makes them especially suitable for use in force fields and for interpreting material properties.

The electron spin density can be described using a four-component spinor, whose four charge density components are directly related to the total electron density, $\rho(\vec{r})$, and the spin magnetization density, $\vec{m}(\vec{r})$. Either the Dirac spinor or the Pauli spin matrices can be used to describe this spinor.^{18–20} To make our method equally applicable for either spin formulation, we use $\rho(\vec{r})$ and $\vec{m}(\vec{r})$ directly. The operator for measuring the spin of electron j is

$$\vec{s}(j) = \hat{x}s_x(j) + \hat{y}s_y(j) + \hat{z}s_z(j) \quad (1)$$

where $s_x(j)$, $s_y(j)$, and $s_z(j)$ are the operators for measuring its spin along the \hat{x} , \hat{y} , and \hat{z} directions, respectively.²¹ Here, we denote the magnitude and direction of a vector by $b = |\vec{b}|$ and $\hat{b} = \vec{b}/b$, respectively. We use δ^{dirac} to denote the Dirac delta function and δ to denote a variational derivative. For a system containing N electrons, the spin magnetization density, $\vec{m}(\vec{r})$, can be computed by summing the spins of all electrons at position \vec{r} and dividing by the spin magnitude of an individual electron. Specifically,

$$\vec{m}(\vec{r}) = 2\langle\Psi| \sum_{j=1}^N (\vec{s}(j)\delta^{\text{dirac}}(\vec{r} - \vec{e}_j))|\Psi\rangle \quad (2)$$

where $\Psi(\{\vec{e}_j\})$ is the multielectronic wave function and $\{\vec{e}_j\}$ are the spatial coordinates of the electrons. The factor of 2 occurs because the spin magnitude of an individual electron is one-half. Spin can be measured along any unit direction \hat{h} , and the value of the spin density projected onto a measurement direction \hat{h} is

$$\rho(\vec{r}, \hat{h}) = (\rho(\vec{r}) + \vec{m}(\vec{r}) \cdot \hat{h})/2 \quad (3)$$

where $\rho(\vec{r})$ and $\vec{m}(\vec{r})$ are expressed in units of electrons per unit volume.²² Collinear magnetism occurs when $\hat{m}(\vec{r})$ is parallel to a global magnetization axis, \hat{h}_{global} , while noncollinear magnetism occurs when it is not. Collinear magnetism has only two independent electron density components, $\rho^\alpha(\vec{r})$ and $\rho^\beta(\vec{r})$, which are the spin density projected onto measurement directions \hat{h}_{global} and $-\hat{h}_{\text{global}}$, respectively.

Since the nonrelativistic and scalar-relativistic Hamiltonians of a chemical system commute with the spin measurement operators, energy eigenstates can be represented as spin eigenstates.²¹ When, however, relativistic spin coupling interactions (e.g., spin-orbit coupling,²³ spin-spin coupling,²⁴ or a nonuniform magnetic field²⁵) are added, the spin measurement operators may no longer commute with the Hamiltonian, leading to energy eigenstates that are not spin eigenstates.^{23–25} We now discuss the theory of spin eigenstates, called pure spin states, for nonperiodic systems.

The multielectronic spin operator, \vec{S} , for a nonperiodic system containing N electrons is²¹

$$\vec{S} = \sum_{j=1}^N \vec{s}(j) \quad (4)$$

Spin eigenstates are described by two quantum numbers and a spin direction.^{21,26} The first spin quantum number (S) satisfies the relationship²¹

$$\vec{S} \cdot \vec{S}|\Psi\rangle = (S(S + 1))|\Psi\rangle \quad (5)$$

Since s_x , s_y , and s_z do not commute, eigenstates along one spin direction are not eigenstates along another spin direction.^{21,26} For the purpose of preparing an orthonormal set of spin eigenstates, the spin direction is conventionally chosen to be \hat{z} .^{21,26} The second spin quantum number (S_z) is the net spin projected onto the spin direction:

$$S_z = \frac{1}{2} \oint \vec{m}(\vec{r}) \cdot \hat{z} d^3\vec{r} \quad (6)$$

where the closed integral symbol indicates integration over a domain without boundary, and since eq 6 is a volume integral this denotes integration over all positions in space.

For a spin eigenstate²¹

$$\vec{S} \cdot \hat{z}|\Psi\rangle = S_z|\Psi\rangle \quad (7)$$

The allowable eigenstates for S_z are

$$S_z \in \{-S, (-S + 1), \dots, S\} \quad (8)$$

giving a spin multiplicity of $2S + 1$.^{21,26} In the absence of spin coupling interactions, these $2S + 1$ states are degenerate. The value of S is (a) 0 for singlet states, (b) 1/2 for doublet states, (c) 1 for a triplet states, etc. The descriptor

$$\Delta_S = \langle\Psi| \vec{S} \cdot \vec{S}|\Psi\rangle - S(S + 1) \quad (9)$$

is useful for quantifying the amount of spin contamination in nonperiodic systems, where $\Delta_S = 0$ for a pure spin state.^{27,28} The situation for periodic materials is more complicated, particularly because the spin structure may not have the periodicity of the ionic structure, leading to phenomena such as spin density waves and frustrated magnetism.^{29–33} For all systems studied in this paper except the single molecule magnet with noncollinear magnetism and spin-orbit coupling, the energy eigenstates should be spin eigenstates.

Several factors need to be considered when computing spin densities using spin density functional theory (DFT).³⁴ For the hypothetically exact exchange-correlation (XC) functional, DFT gives the wave function of a fictitious noninteracting system whose energy, $\rho(\vec{r})$, and $\vec{m}(\vec{r})$ are the same as in the real interacting system.³⁵ Since ASMs are directly derived from $\rho(\vec{r})$ and $\vec{m}(\vec{r})$, the DFT wave function computed from the exact XC functional would give the same ASMs as the wave function of the real interacting system. The DFT wave function of the noninteracting system need not be a spin eigenstate, even though it formally reproduces the spin densities of the real interacting system.^{34,36} In practice, systems for which the DFT wave function has a small (large) amount of spin contamination are easier (harder) to accurately model. For this reason, we present and discuss Δ_S values for the nonperiodic DFT wave functions in the Results below. One must be particularly careful when using DFT to compute the spin densities of multireference systems

like open shell singlets and states for which $|S_z| < S$, since the symmetry of the spin density may not be well-represented by a single Slater determinant.^{37–40} Extensions of DFT using a linear combination of Slater determinants have been developed to handle these difficult cases and to approximately treat excited states.^{37,41–44}

In this article, we focus mainly on nonrelativistic and scalar-relativistic Hamiltonians where energy eigenstates should in principle be representable as spin eigenstates. We study a variety of periodic and nonperiodic collinear magnetic systems using traditional DFT methods, including a DFT+U and range-separated hybrid study of the charge-spin ordering in bulk magnetite. We use coupled-cluster expansions to study various spin states of ozone. Coupled cluster methods that give formally exact expansions of the system's wave function within the nonrelativistic and scalar-relativistic Born–Oppenheimer approximations are a useful and powerful tool for modeling multireference systems.^{45–49} In addition, a noncollinear single molecule magnet is studied using noncollinear DFT including spin–orbit coupling interactions. Several population analysis methods are applied to each of these systems. In agreement with prior literature, we find the choice of population analysis method has a smaller effect on the computed ASMs than on the computed NACs.⁵⁰ We find the DDEC method described in this article gives reliable NACs and ASMs for all the materials studied.

The rest of this paper is organized as follows. Section 2 describes a theory for partitioning the spin density among atoms in a material to give ASMs that simultaneously reproduce the chemical states of atoms and $\bar{B}^{\text{spin}}(\vec{r})$ outside the material's electron distribution. Section 2 also describes the theory of choosing an appropriate XC functional for modeling different types of noncollinear magnetism. Section 3 describes the computational methods used to compute optimized geometries, electron density distributions, NACs, and ASMs. In section 4, we use these methods to study a variety of periodic and nonperiodic materials with collinear and noncollinear magnetism. We describe the main conclusions of this work in section 5. The Appendix proves that the optimization functional for computing ASMs converges to a unique solution and also contains algorithms for computing a function and constant used in the iterative solution method.

2. THEORY

2.1. Atom-in-Molecules (AIM) Formalism. This section describes universal equations for partitioning the total spin density among atoms in a material, for computing ASMs from atomic spin density distributions, for expressing $\bar{B}^{\text{spin}}(\vec{r})$ as a distributed multipole expansion, and for quantifying how accurately a set of ASMs reproduce $\bar{B}^{\text{spin}}(\vec{r})$. These equations apply to all atoms-in-molecule methods. We begin by defining an arbitrary material as a set of atoms $\{A\}$ at positions $\{\vec{R}_A\}$, in a reference unit cell, \mathbf{U} .¹⁴ For a nonperiodic system (e.g., a molecule), \mathbf{U} is any parallelepiped enclosing the entire electron distribution. Using this notation, $\vec{m}(\vec{r})$ can be written as a sum of atomic spin magnetization densities

$$\vec{m}(\vec{r}) = \vec{\Delta}(\vec{r}) + \sum_{k,A} \vec{m}_A(\vec{r}_A) \quad (10)$$

where the constraint

$$\vec{\Delta}(\vec{r}) = \vec{0} \quad (11)$$

will be imposed below using a Lagrange multiplier. Throughout this article, we use the short-hand notation

$$\sum_{k,A} = \sum_{k_1} \sum_{k_2} \sum_{k_3} \sum_A \quad (12)$$

The reference unit cell has $k_1 = k_2 = k_3 = 0$ and summation over A means summation over all atoms in this unit cell. For a periodic direction, k_i ranges over all integers with the associated lattice vector \vec{v}_i . For a nonperiodic direction, $k_i = 0$ and \vec{v}_i is the corresponding edge of \mathbf{U} . Using this notation, the vector and distance relative to atom A are given by

$$\vec{r}_A = \vec{r} - k_1 \vec{v}_1 - k_2 \vec{v}_2 - k_3 \vec{v}_3 - \vec{R}_A \quad (13)$$

and $r_A = |\vec{r}_A|$.¹⁴ Analogous to eq 3, the value of the atomic spin density projected onto a measurement direction \hat{h} is

$$\rho_A(\vec{r}_A, \hat{h}) = (\rho_A(\vec{r}_A) + \vec{m}_A(\vec{r}_A) \cdot \hat{h})/2 \quad (14)$$

The atomic electron distribution, $\rho_A(\vec{r}_A)$, is computed via an atoms-in-molecule (AIM) formalism such as Bader,⁴⁵ DDEC,¹⁴ Hirshfeld,⁶ iterative Hirshfeld,¹² or iterative Stockholder atom⁵¹ analysis.

The ASM is the vector \vec{M}_A computed by integrating $\vec{m}_A(\vec{r}_A)$,

$$\vec{M}_A = \oint \vec{m}_A(\vec{r}_A) d^3\vec{r}_A \quad (15)$$

For collinear magnetism, the ASM can be equivalently expressed as the scalar projection of \vec{M}_A onto the global spin quantization axis:

$$T_A = \vec{M}_A \cdot \hat{h}_{\text{global}} \quad (16)$$

The unit cell spin moment, \vec{M} , is the sum of individual ASMs,

$$\vec{M} = \sum_A \vec{M}_A \quad (17)$$

For collinear magnetism, the unit cell spin moment can be equivalently expressed as the scalar projection of \vec{M} onto the global spin quantization axis:

$$T = \vec{M} \cdot \hat{h}_{\text{global}} = \sum_A T_A \quad (18)$$

Since the spin-derived magnetic moment of a spin-up electron is $g_e \mu_B/2$ where $g_e \approx -2.002319$ is the electron's g factor and μ_B is the Bohr magneton,⁵² all spin-derived magnetic moments in this paper are reported in units of $g_e \mu_B/2$. (Because of its negative charge, the electron's spin and magnetic moments point in opposite directions, causing g_e to be negative.) In these units, each electron contributes unit magnitude. For collinear systems, T_A is the number of spin-up minus spin-down electrons assigned to atom A , and T is the number of spin-up minus spin-down electrons in the unit cell.

Subject to the above constraints, the key issue is how to optimize $\{\vec{m}_A(\vec{r}_A)\}$ to simultaneously reproduce $\bar{B}^{\text{spin}}(\vec{r})$ and the chemical states of atoms. An atomic expansion of $\bar{B}^{\text{spin}}(\vec{r})$ has the form

$$\bar{B}^{\text{spin}}(\vec{r}) = \sum_{k,A} \bar{B}_A^{\text{spin}}(\vec{r}_A) \quad (19)$$

We assume the wave function and other properties are not time-dependent. Setting the density of point magnetic dipoles equal to $g_e \mu_B \vec{m}_A(\vec{r}_A)/2$ and integrating the classical expression for the

magnetic field of a point dipole^{52,53} over \vec{r}'_A gives

$$\begin{aligned} \vec{B}_A^{\text{spin}}(\vec{r}_A) &= \frac{\mu_0}{4\pi} \left(\frac{g_e \mu_B}{2} \right) \oint \left(\frac{3(\vec{r}_A - \vec{r}'_A)(\vec{m}_A(\vec{r}'_A) \cdot (\vec{r}_A - \vec{r}'_A))}{|\vec{r}_A - \vec{r}'_A|^5} \right. \\ &\quad \left. - \frac{\vec{m}_A(\vec{r}'_A)}{|\vec{r}_A - \vec{r}'_A|^3} \right) d^3 r'_A \end{aligned} \quad (20)$$

One can rewrite eq 20 with the help of a multipolar expansion as

$$\begin{aligned} \vec{B}_A^{\text{spin}}(\vec{r}_A) &= \frac{\mu_0}{4\pi} \left(\frac{g_e \mu_B}{2} \right) \left(\frac{3\vec{r}_A (\vec{M}_A \cdot \vec{r}_A)}{(r_A)^5} - \frac{\vec{M}_A}{(r_A)^3} \right) \\ &\quad + Q_A + P_A \end{aligned} \quad (21)$$

where magnetic dipoles are the leading term. The terms Q_A and P_A are due to higher order magnetic multipoles and electron cloud penetration effects, respectively. The penetration term P_A is essentially zero for r_A values greater than a critical radius where $\rho_A(\vec{r}_A)$ is negligible. For the development of force fields, it is convenient to choose values for $\{M_A\}$ such that Q_A is small for large distances where P_A is negligible. Under these conditions, $\vec{B}^{\text{spin}}(\vec{r})$ can be accurately reproduced using just the $\{M_A\}$ values, which facilitates constructing a force field to reproduce $\vec{B}^{\text{spin}}(\vec{r})$. Due to symmetry, Q_A is zero when $\vec{m}_A(\vec{r}_A)$ is spherically symmetric, that is, when $\vec{m}_A(\vec{r}_A) = \vec{m}_A^{\text{avg}}(r_A)$. Thus, to simultaneously reproduce an atom's chemical state and $\vec{B}_A^{\text{spin}}(\vec{r}_A)$ for large r_A values, $\vec{m}_A(\vec{r}_A)$ should be optimized to resemble $\vec{m}_A^{\text{avg}}(r_A)$ subject to constraint 26, where the spherical average $g_A^{\text{avg}}(r_A)$ of a scalar or vector field $g_A(\vec{r}_A)$ is defined by

$$g_A^{\text{avg}}(r_A) = \frac{1}{4\pi(r_A)^2} \oint g_A(\vec{r}'_A) \delta^{\text{dirac}}(r_A - r'_A) d^3 r'_A \quad (22)$$

The potential energy of a charge, q , and magnetic dipole, \vec{M} , placed in an electrostatic potential $V(\vec{r})$ and magnetostatic field $\vec{B}(\vec{r})$ is

$$E = qV(\vec{r}) - \vec{M} \cdot \vec{B}(\vec{r}) \quad (23)$$

Thus, to obtain the most accurate energies in atomistic simulations, NACs and ASMs should be optimized to reproduce $V(\vec{r})$ and $\vec{B}^{\text{spin}}(\vec{r})$ outside a material's van der Waals surface. For time-independent fields, such as those considered here, $V(\vec{r})$ is determined up to an additive constant, and $\vec{B}^{\text{spin}}(\vec{r})$ is completely determined (i.e., gauge invariant). Earlier papers described a method for computing the root mean squared error (RMSE) between $V(\vec{r})$ and the electrostatic potential generated by a set of NACs in a way that automatically corrects for the arbitrary constant in $V(\vec{r})$.^{16,54} We now describe a method for quantifying the accuracy of ASMs for reproducing $\vec{B}^{\text{spin}}(\vec{r})$. The accuracy of each ASM model for reproducing $\vec{B}^{\text{spin}}(\vec{r})$ was quantified by the mean absolute error (MAE):

$$\text{MAE} = \sum_{\vec{p}} |(\vec{B}^{\text{spin}}(\vec{p})|_{\text{ASM}} - \vec{B}^{\text{spin}}(\vec{p})|_{\text{spindensity}})|u(\vec{p})| / \sum_{\vec{p}} u(\vec{p}) \quad (24)$$

where $\vec{B}^{\text{spin}}(\vec{p})|_{\text{ASM}}$ and $\vec{B}^{\text{spin}}(\vec{p})|_{\text{spindensity}}$ are the values at grid point \vec{p} due to the ASM's and ab initio spin density, respectively. The finite element volume, $u(\vec{p})$, for each grid point makes the summations in eq 24 equivalent to corresponding integrals over volume. $\vec{B}^{\text{spin}}(\vec{p})|_{\text{spindensity}}$ was computed by numerically

integrating

$$\vec{B}^{\text{spin}}(\vec{p}) = \frac{\mu_0}{4\pi} \left(\frac{g_e \mu_B}{2} \right) \oint \left(\frac{3(\vec{p} - \vec{r}')(\vec{m}(\vec{r}') \cdot (\vec{p} - \vec{r}'))}{|\vec{p} - \vec{r}'|^5} - \frac{\vec{m}(\vec{r}')}{|\vec{p} - \vec{r}'|^3} \right) d^3 r' \quad (25)$$

We found that integrating \vec{r}' over grid points inside the surface defined by $2.4 \times \text{vdW}$ radii enclosed essentially all of the spin density. (The vdW radii listed in the Supporting Information of Watanabe et al. were used.¹⁶) To avoid division by $|\vec{p} - \vec{r}'| = 0$ in eq 25, the set of grid points, $\{\vec{p}\}$, used to compute the MAE were distributed in the volume between surfaces defined by $3 \times$ and $4 \times$ vdW radii. The relative mean absolute error (RMAE) was then defined as the MAE for an ASM model divided by the MAE when all ASMs were set to zero.

2.2. New Optimization Functional for Computing ASMs.

Now that general equations applicable to all AIM methods have been described in the previous section, we describe strategies for optimizing ASMs using AIM methods like DDEC that simultaneously optimize the $\rho_A(\vec{r}_A)$ distributions to be close to spherical symmetry and to resemble the electron distributions of isolated reference atoms.¹⁴ To be chemically meaningful, the spin projected electron density $\rho_A(\vec{r}_A, \hat{h})$ must always be non-negative, which requires

$$\kappa_A(\vec{r}_A) = \rho_A(\vec{r}_A) - m_A(\vec{r}_A) \geq 0 \quad (26)$$

Equation 26 confines $m_A(\vec{r}_A)$ to the region of space where $\rho_A(\vec{r}_A)$ is non-negligible. Since DDEC $\rho_A(\vec{r}_A)$ distributions decay approximately exponentially with increasing r_A ,¹⁴ $m_A(\vec{r}_A)$ becomes negligible for sufficiently large r_A . In the remainder of this section, a suitable functional for optimizing $m_A(\vec{r}_A)$ subject to constraint 26 is derived.

One of the simplest optimization functionals would be to minimize the weighted sum of $|\vec{m}_A(\vec{r}_A) - \vec{m}_A^{\text{avg}}(r_A)|^2$ subject to constraints 11 and 26, which is denoted LSF (least-squares fitting) in the Results below. After programming this functional as described in the Supporting Information, we found that it performs poorly for dense systems containing adjacent magnetic atoms. For example, LSF requires 2094 iterations to converge the ASMs for solid Fe₃Si. Such a large number of iterations usually indicates a nearly flat optimization landscape. Nearly flat landscapes are problematic because the location of the minimum is unduly sensitive to small perturbations. As shown in the Results below, LSF gives the wrong sign for the ASM of Si in solid Fe₃Si. Earlier, we encountered analogous problems when $\rho_A(\vec{r}_A)$ was optimized to be as close to $\rho_A^{\text{avg}}(r_A)$ as possible.¹⁴ In that case, the remedy was an optimization functional that minimizes the sum of $\rho_A(\vec{r}_A) \ln(\rho_A(\vec{r}_A)/w_A(r_A))$ where $w_A(r_A)$ is a weighted geometric average of a reference density and $\rho_A^{\text{avg}}(r_A)$ subject to the constraint that $w_A(r_A)$ does not become too diffuse.¹⁴ A functional of this type minimizes the information distance between $\rho_A(\vec{r}_A)$ and $w_A(r_A)$ subject to the applied constraints. We found a similar strategy works for optimizing the atomic spin densities and reduces the number of iterations to approximately 10—a 2 orders of magnitude improvement. As shown in the Results below, this strategy gives accurate ASMs for solid Fe₃Si and other materials.

Specifically, we construct an optimization functional for the spin density projected onto the unit direction, \hat{h} , by minimizing the information distance $\rho_A(\vec{r}_A, \hat{h}) \ln(\rho_A(\vec{r}_A, \hat{h})/w_A^{\text{spin}}(\vec{r}_A, \hat{h}))$ between $\rho_A(\vec{r}_A, \hat{h})$ and $w_A^{\text{spin}}(\vec{r}_A, \hat{h})$ subject to constraints 11 and 26. We found that excellent performance is obtained when

$w_A^{\text{spin}}(\vec{r}_A, \hat{h})$ is a weighted geometric average between proportional spin partitioning and $\rho_A^{\text{avg}}(\vec{r}_A, \hat{h})$

$$w_A^{\text{spin}}(\vec{r}_A, \hat{h}) = (\rho_A^0(\vec{r}_A, \hat{h}))^{\chi_{\text{spin}}} (\rho_A^{\text{avg}}(\vec{r}_A, \hat{h}))^{1 - \chi_{\text{spin}}} \quad (27)$$

where proportional spin partitioning is defined by

$$\vec{m}_A^0(\vec{r}_A) = \vec{m}(\vec{r}) \rho_A(\vec{r}_A) / \rho(\vec{r}) \quad (28)$$

This has a form analogous to the weight used in DDEC charge partitioning.¹⁴ Reasonable values of χ_{spin} are in the range $0 < \chi_{\text{spin}} \leq 1$. In the Results below, we study $\chi_{\text{spin}} = 3/14$ (~ 0.21), 0.5, and 1. $\chi_{\text{spin}} = 1$ gives proportional spin partitioning. To accommodate both collinear and noncollinear magnetism, integration is performed over all possible choices of \hat{h} . This gives the following optimization functional:

$$H = \sum_A \left(\int_{\omega} \int \rho_A(\vec{r}_A, \hat{h}) \ln \left(\frac{\rho_A(\vec{r}_A, \hat{h})}{w_A^{\text{spin}}(\vec{r}_A, \hat{h})} \right) d^3 \vec{r}_A d^2 \omega \right. \\ \left. - \int \nu_A(\vec{r}_A) \kappa_A(\vec{r}_A) d^3 \vec{r}_A \right) + \int_U \vec{\Lambda}(\vec{r}) \vec{\Delta}(\vec{r}) d^3 \vec{r} \quad (29)$$

where the Lagrange multipliers $\nu(\vec{r}_A) \geq 0$ and $\vec{\Lambda}(\vec{r})$ are used to enforce constraints 26 and 11, respectively. Integration over ω means integration over all possible choices for \hat{h} , because ω is the unit sphere surface comprised of all possible end points for \hat{h} . Minimization occurs when

$$\delta H = \sum_A \int \frac{\partial H}{\partial \vec{m}_A(\vec{r}_A)} \cdot \delta \vec{m}_A(\vec{r}_A) d^3 \vec{r}_A = 0 \quad (30)$$

for arbitrary $\delta \vec{m}_A(\vec{r}_A)$, which is true when

$$\frac{\partial H}{\partial \vec{m}_A(\vec{r}_A)} = \frac{1}{2} \int_{\omega} \ln \left(\frac{\rho_A(\vec{r}_A, \hat{h})}{w_A^{\text{spin}}(\vec{r}_A, \hat{h})} \right) \hat{h} d^2 \omega - \vec{\Lambda}(\vec{r}) \\ + \nu_A(\vec{r}_A) \hat{m}_A(\vec{r}_A) \quad (31)$$

equals zero, where the vector partial derivative is defined as

$$\frac{\partial}{\partial \vec{b}} = \hat{x} \frac{\partial}{\partial b_x} + \hat{y} \frac{\partial}{\partial b_y} + \hat{z} \frac{\partial}{\partial b_z} \quad (32)$$

To evaluate the integral in eq 31, we first define the following function of τ with $0 \leq \tau \leq 1$:

$$\xi(\tau) \hat{\tau} = \int_{\omega} \ln(1 + \vec{\tau} \cdot \hat{h}) \hat{h} d^2 \omega \quad (33)$$

By symmetry, the integral on the right-hand side must be parallel to $\hat{\tau}$. Without a loss of generality, we choose a set of spherical coordinates having $\hat{z} = \hat{\tau}$, and then rewrite eq 33 as

$$\xi(\tau) = \int_0^\pi \int_0^{2\pi} \cos(\phi) \ln(1 + \tau \cos(\phi)) \sin(\phi) d\theta d\phi \quad (34)$$

Analytic evaluation of this integral gives

$$\xi(\tau) = \pi \left((\tau^{-2} - 1) \ln \left(\frac{1 - \tau}{1 + \tau} \right) + \frac{2}{\tau} \right) \quad (35)$$

Note that $\xi(0) = 0$ and $\xi(1) = 2\pi$. After defining the functions

$$\vec{\vartheta}(a, \vec{b}) = \frac{\hat{b}}{2} \xi \left(\frac{b}{a} \right) \quad (36)$$

$$\vec{L}_A(\vec{r}_A) = \nu_A(\vec{r}_A) \hat{m}_A(\vec{r}_A) + \vec{\vartheta}(\rho_A(\vec{r}_A), \vec{m}_A(\vec{r}_A)) \quad (37)$$

$$\vec{\eta}_A(\vec{r}_A) = (1 - \chi_{\text{spin}}) \vec{\vartheta}(\rho_A^{\text{avg}}(\vec{r}_A), \vec{m}_A^{\text{avg}}(\vec{r}_A)) \\ + \chi_{\text{spin}} \vec{\vartheta}(\rho_A(\vec{r}_A), \vec{m}_A^0(\vec{r}_A)) \quad (38)$$

$$\vec{\Lambda}_A(\vec{r}_A) = \vec{L}_A(\vec{r}_A) - \vec{\eta}_A(\vec{r}_A) \quad (39)$$

equations 14, 31, 33, and 36–39 can be combined to give the solution of the optimization problem

$$\vec{\Lambda}_A(\vec{r}_A) = \vec{\Lambda}(\vec{r}) \quad (40)$$

A proof that $\{\vec{m}_A(\vec{r}_A)\}$ are uniquely determined because H has only one minimum is given in the Appendix.

2.3. DFT Functionals for Noncollinear Magnetism. The accuracy of computed ASMs depends on the accuracy of the XC functional used to compute $\rho(\vec{r})$ and $\vec{m}(\vec{r})$. In noncollinear DFT, the total XC energy is formally a unique functional of the four spinor components at all positions in space:^{25,35}

$$E_{\text{XC}}^{\text{non-collinear}} = E_{\text{XC}}^{\text{four}}(\{\rho(\vec{r}), \vec{m}(\vec{r})\}) \quad (41)$$

On the other hand, the total XC energy for collinear magnetism is formally a unique functional of two electron density components, $\rho^\alpha(\vec{r})$ and $\rho^\beta(\vec{r})$, at all positions in space:²⁵

$$E_{\text{XC}}^{\text{collinear}}(\{\rho^\alpha(\vec{r}), \rho^\beta(\vec{r})\}) \quad (42)$$

All derivatives of the electron and spin densities are already included in functionals 41 and 42, because a function's derivatives are computed from values of that function at different positions. Density functionals originally developed for collinear magnetism are commonly extended to noncollinear magnetism by replacing $\rho^\alpha(\vec{r})$ and $\rho^\beta(\vec{r})$ in eq 42 with the spin densities projected onto the local spin magnetization direction:^{25,35}

$$\rho^{(+)}(\vec{r}) = \rho(\vec{r}, +\hat{m}(\vec{r})) = (\rho(\vec{r}) + m(\vec{r}))/2 \quad (43)$$

$$\rho^{(-)}(\vec{r}) = \rho(\vec{r}, -\hat{m}(\vec{r})) = (\rho(\vec{r}) - m(\vec{r}))/2 \quad (44)$$

to give an XC functional that depends only on $\rho(\vec{r})$ and $m(\vec{r})$:

$$E_{\text{XC}}^{\text{two}}(\{\rho(\vec{r}), m(\vec{r})\}) \quad (45)$$

Functional 45, which can be conveniently applied to both collinear and noncollinear magnetism, is formally exact for collinear magnetism because it involves two spinor components. Local spin density approximations (LSDA) and generalized gradient approximations (GGA) are explicit analytic functionals of form 42 that can readily be converted to form 45 using eqs 43 and 44.

To quantify the inherent limitations of $E_{\text{XC}}^{\text{two}}$ for modeling noncollinear magnetism, we consider its ability to reproduce the derivatives

$$\nu_{\text{XC}}(\vec{r}) = \frac{\partial E_{\text{XC}}}{\partial \rho(\vec{r})} \quad (46)$$

$$\vec{B}_{XC}(\vec{r}) = \frac{\partial E_{XC}}{\partial \vec{m}(\vec{r})} \quad (47)$$

where $\nu_{XC}(\vec{r})$ and $\vec{B}_{XC}(\vec{r})$ are called the XC potential and XC magnetic field, respectively.^{55,56} Inserting eq 45 into 47 and using the chain rule of differentiation gives

$$\vec{B}_{XC}^{two}(\vec{r}) = \frac{\partial m(\vec{r})}{\partial \vec{m}(\vec{r})} \frac{\partial E_{XC}^{two}(\{\rho(\vec{r}), m(\vec{r})\})}{\partial m(\vec{r})} \quad (48)$$

Since $\partial m(\vec{r})/\partial \vec{m}(\vec{r}) = \hat{m}(\vec{r})$, $\vec{B}_{XC}^{two}(\vec{r})$ is always parallel to $\hat{m}(\vec{r})$:

$$\vec{B}_{XC}^{two}(\vec{r}) = \hat{m}(\vec{r}) \frac{\partial E_{XC}^{two}(\{\rho(\vec{r}), m(\vec{r})\})}{\partial m(\vec{r})} \quad (49)$$

Capelle et al. showed that the exact $\vec{B}_{XC}(\vec{r})$ is not parallel to $\vec{m}(\vec{r})$ if $\vec{\nabla} \cdot [J^{KS}(\vec{r}) - J(\vec{r})] \neq 0$, where $J^{KS}(\vec{r})$ and $J(\vec{r})$ are the spin-current tensors of the DFT noninteracting system and real interacting system, respectively.⁵⁵ Thus, eq 49 describes a fundamental limitation of all XC functionals that depend only on $\{\rho(\vec{r}), m(\vec{r})\}$. This limitation was previously reported for LSDA and GGA functionals.^{55,56}

What is missing from E_{XC}^{two} that forces $\vec{B}_{XC}^{two}(\vec{r})$ to be parallel to $\hat{m}(\vec{r})$? The functional E_{XC}^{two} does not include any dependence on $\hat{m}(\vec{r})$. These directional changes can be quantified by the non-negative descriptors

$$f^{dir}(\vec{r}) = m(\vec{r})^2 \sum_{i=1}^3 (\nabla_i \hat{m}(\vec{r}) \cdot \nabla_i \hat{m}(\vec{r})) \geq 0 \quad (50)$$

$$\Xi^{dir} = \int_{\Omega} f^{dir}(\vec{r}) d^3\vec{r} \geq 0 \quad (51)$$

Analogous descriptors can be defined for the changes in $m(\vec{r})$

$$f^{mag}(\vec{r}) = \vec{\nabla} m(\vec{r}) \cdot \vec{\nabla} m(\vec{r}) \geq 0 \quad (52)$$

$$\Xi^{mag} = \int_{\Omega} f^{mag}(\vec{r}) d^3\vec{r} \geq 0 \quad (53)$$

Non-negative measures of all gradient changes in $\vec{m}(\vec{r})$ are

$$f^{tot}(\vec{r}) = \sum_{i=1}^3 (\nabla_i \vec{m}(\vec{r}) \cdot \nabla_i \vec{m}(\vec{r})) \\ = f^{dir}(\vec{r}) + f^{mag}(\vec{r}) \geq 0 \quad (54)$$

$$\Xi^{tot} = \Xi^{mag} + \Xi^{dir} \geq 0 \quad (55)$$

Clearly, $\Xi^{tot} = 0$ for nonmagnetic systems, and $\Xi^{tot} > 0$ for magnetic systems. Similarly, $\Xi^{dir} = 0$ for collinear magnetism, and $\Xi^{dir} > 0$ for noncollinear magnetism. Thus, it follows that

$$\lim_{\Xi^{dir} \rightarrow 0} E_{XC}^{four} = E_{XC}^{two} \quad (56)$$

In practice, we expect E_{XC}^{two} to provide a reasonable approximation to the XC energy of a system's ground state when most of the $\vec{m}(\vec{r})$ changes are due to $m(\vec{r})$ changes, i.e., when $\Xi^{dir} \ll \Xi^{mag}$. For the ground states of many noncollinear magnetic materials, $\hat{m}(\vec{r})$ changes significantly only in regions where $m(\vec{r})$ is small,^{22,57–59} leading to $\Xi^{dir} \ll \Xi^{mag}$ and $E_{XC}^{non-collinear} \approx E_{XC}^{two}$.

For example, consider the spin spiral ground state that has been experimentally observed in γ -Fe.⁶⁰ In this state, the ASMs exhibit a combined rotation-translation symmetry group.⁵⁹ \hat{M}_A is significantly different for different Fe atoms, even though M_A is similar.⁵⁹ Furthermore, $\hat{m}(\vec{r})$ is approximately constant inside a

single atomic volume and changes only in regions of space where $m(\vec{r})$ is small.⁵⁹ From this, we infer $\Xi^{dir} \ll \Xi^{mag}$. Accordingly, computations showed E_{XC}^{two} using the GGA approximation correctly reproduces the spin spiral ground state of γ -Fe for an appropriate range of lattice constants.⁵⁹ In summary, we derived the conditions under which the XC functional is accurately modeled by E_{XC}^{two} . This is used to apply the PW91 functional to noncollinear magnetism in section 4.3 below.

3. METHODS

3.1. Iterative Solution Algorithm for ASMs. We showed above that the task of assigning ASMs can be formulated in terms of optimizing the functional H defined in eq 29. We now describe an iterative algorithm that efficiently converges to H 's unique minimum. Each iteration involves three loops. The first loop runs over $\{A, \vec{r}_A\}$, where all grid points having $r_A \leq r_{cutoff}$ are included. A cutoff of 4 Å was used in all calculations reported below. This loop generates an updated and unstored estimate of $\vec{m}_A(\vec{r}_A)$ that is used to accumulate sums stored for use in the second and third loops. In the first iteration, proportional partitioning is used:

$$\vec{m}_A(\vec{r}_A)|_1 = \vec{m}_A^0(\vec{r}_A) \quad (57)$$

$$\nu_A(\vec{r}_A)|_1 = 0 \quad (58)$$

For reasons that will be explained below, in subsequent iterations, an estimate for $\vec{L}_A(\vec{r}_A)$ is computed by

$$\vec{L}_A(\vec{r}_A) = \vec{Y}(\vec{r}) - \vec{Q}(\vec{r}) \\ + c \frac{(\vec{m}(\vec{r}) - \vec{m}^{trial}(\vec{r}))}{\rho(\vec{r})} + \vec{\eta}_A(\vec{r}_A) \quad (59)$$

The optimal value $c = \pi$ is derived in the Appendix. The magnitude $L_A(\vec{r}_A)$ was used to update the estimate for $m_A(\vec{r}_A)$ according to two cases. Case 1: if $L_A(\vec{r}_A) > \pi$, then $m_A(\vec{r}_A) = \rho_A(\vec{r}_A)$ and $\nu_A(\vec{r}_A) = L_A(\vec{r}_A) - \pi > 0$. Case 2: if $L_A(\vec{r}_A) \leq \pi$, $m_A(\vec{r}_A) = \rho_A(\vec{r}_A) \xi^{inv}(2L_A(\vec{r}_A))$ and $\nu_A(\vec{r}_A) = 0$, where the algorithm for computing ξ^{inv} is given in the Appendix. These two cases ensure that eq 26 is satisfied for all iterations and $\nu_A(\vec{r}_A) \geq 0$. The updated magnetization is given by

$$\vec{m}_A(\vec{r}_A) = m_A(\vec{r}_A) \hat{L}_A(\vec{r}_A) \quad (60)$$

During this loop, the sum of valid grid points for each r_A is accumulated and stored, as well as the accumulated sum of $\vec{m}_A(\vec{r}_A)$ for each r_A value. Finally, the following sums are also accumulated and stored

$$\vec{Y}(\vec{r}) = \sum_{k,A} \vec{L}_A(\vec{r}_A) \frac{\rho_A(\vec{r}_A)}{\rho(\vec{r})} \quad (61)$$

$$\vec{m}^{trial}(\vec{r}) = \sum_{k,A} \vec{m}_A(\vec{r}_A) \quad (62)$$

The second loop runs over $\{A, r_A\}$. This loop computes and stores $\vec{m}_A^{avg}(r_A)$, $m_A^{avg}(r_A)$, and M_A from the sums accumulated in the first loop. The third loop runs over $\{A, \vec{r}_A\}$ like the first loop. This loop accumulates and stores

$$\vec{Q}(\vec{r}) = \sum_{k,A} \left(\vec{\eta}_A(\vec{r}_A) \frac{\rho_A(\vec{r}_A)}{\rho(\vec{r})} \right) \quad (63)$$

After the third loop, the program starts the next iteration by going back to the first loop. At least five iterations were performed in all cases. Additional iterations were performed until all components in $\{\bar{M}_A\}$ changed by less than 5×10^{-5} .

We now show this iterator robustly converges to H 's global minimum. Upon convergence, multiplying eq 59 by $\rho_A(\vec{r}_A)/\rho(\vec{r})$ and summing over $\{k_1, k_2, k_3, A\}$ gives

$$\vec{0} = c(\vec{m}(\vec{r}) - \vec{m}^{\text{trial}}(\vec{r}))/\rho(\vec{r}) \quad (64)$$

which proves constraint 11 is satisfied. Therefore, at convergence, eq 59 gives

$$\bar{\Lambda}_A(\vec{r}_A) = \bar{L}_A(\vec{r}_A) - \vec{\eta}_A(\vec{r}_A) = \vec{Y}(\vec{r}) - \vec{Q}(\vec{r}) = \bar{\Lambda}(\vec{r}) \quad (65)$$

which is the required solution. The iterative scheme converges $\bar{\Lambda}_A(\vec{r}_A)$ to a function $\bar{\Lambda}(\vec{r})$ that is the same for all atoms using an averaging mechanism. Specifically, in each iteration, the updated estimate for the function $\bar{\Lambda}(\vec{r}) = \vec{Y}(\vec{r}) - \vec{Q}(\vec{r})$ is constructed by averaging the $\bar{\Lambda}_A(\vec{r}_A)$ values over all atoms. Using eq 59, this $\bar{\Lambda}(\vec{r})$ estimate is used in the subsequent iteration to compute a new estimate for $\bar{\Lambda}_A(\vec{r}_A)$. This causes the estimated $\bar{\Lambda}_A(\vec{r}_A)$ for all atoms to rapidly converge toward an average value, $\bar{\Lambda}(\vec{r})$, thereby solving the optimization problem. When using $\chi_{\text{spin}} = 1/2$ or $3/14$, we have not encountered any system for which more than 10 and 20 iterations, respectively, were required to converge $\{\bar{M}_A\}$ to within 5×10^{-5} .

Our implementation is computationally efficient. For materials containing several hundred atoms, a large amount of memory would be required to store the complete set of functions of \vec{r}_A . The above approach minimizes overall memory requirements by only requiring functions of \vec{r}_A to be stored for the current position. This allows systems containing hundreds of atoms in the unit cell to be analyzed on a typical computer processor with a few gigabytes of memory. Since only complete functions of \vec{r} and r_A need to be stored, the overall memory requirements scale linearly with the number of grid points (i.e., volume) in the unit cell. The use of a cutoff radius causes the overall computational time to scale linearly with the number of atoms in the unit cell. In summary, both the computational time and the memory requirements scale linearly with increasing system size, which makes the algorithm efficient for both small and large systems. These same properties hold for the DDEC charge analysis.¹⁴ For collinear magnetism, all $\vec{m}_A(\vec{r}_A)$ are parallel to a global magnetization axis. In this case, vectors can be replaced with their scalar projection onto the global magnetization axis. It is therefore only necessary to compute and store one-third as many components as in the noncollinear case. For collinear magnetism, our implementation automatically avoids the computation and storage of all zero-valued spin components. Except for this simplification, all other details are the same for the collinear and noncollinear cases.

3.2. Geometry and Electron Density Generation. All periodic DFT calculations were performed in the Vienna Ab Initio Simulation Package (VASP)^{61,62} using the projector augmented wave (PAW) method⁶³ and a 400 eV cutoff. For all periodic systems, the product of the number of k points and the unit cell volume exceeded 4000 \AA^3 , except for the HSE06 calculation of magnetite, which was performed only at the Γ point. All non-periodic calculations were performed using Gaussian software.⁶⁴ The XC functionals, basis sets, and additional details are described in the Results section below. The PW91-optimized

geometry⁶⁵ for the $\text{Cu}_3(\text{BTC})_2$ metal organic framework (MOF) and experimental Fe_3Si crystal geometry⁶⁶ were taken from the literature. The remaining geometries were optimized in VASP or Gaussian. In VASP, geometries were optimized to give atomic forces $<0.03 \text{ eV/\AA}$. In Gaussian, geometries were optimized to give atomic forces and displacements smaller than 0.0025 au and 0.01 au, respectively.

Core–valence XC interactions are critical for describing the properties of magnetic materials.⁶⁷ The PAW method has many advantages, especially for studying magnetic materials, because it is an all-electron frozen core method that explicitly includes core–valence XC interactions.⁶³ The PAW method has higher accuracy and similar computational cost to pseudopotential methods;⁶³ however, pseudopotentials that use partial core corrections to describe valence–core XC interactions can also give accurate results including magnetic properties.⁶⁷ For calculations employing linear combination of atomic orbital (LCAO) basis sets, relativistic effective core potentials (RECPs) that include outer-core shell electrons in the valence space can give accurate magnetic properties because they describe the most important valence–core XC interactions.⁶⁸ This is more important for s -, d -, and f -block elements than for p -block elements.^{67,68} We used the PAW method for all periodic calculations in this paper. For nonperiodic calculations, we used all-electron basis sets on lighter atoms and appropriate RECPs for heavier atoms. The specific basis sets are listed in the Results below.

DDEC atomic charge distributions were computed using the DDEC/c2 method of Manz and Sholl.¹⁴ Spherical averaging was performed using 75 radial shells evenly spaced between $r_A = 0$ and the cutoff radius of 4 \AA . To make sure all electrons were included, core electrons replaced by a pseudopotential or effective core potential during a quantum chemistry calculation were automatically reinserted by the program at the beginning of DDEC analysis. (The program stores core electron reference densities for all chemical elements.) The speed of core charge fitting was improved by applying the constraint

$$w_A^{\text{core}}(r_A) = \min(\rho_A^{\text{core}}(r_A), \rho_A^{\text{core}}(r_A - \Delta r_A) e^{-2\Delta r_A/\text{bohr}}) \quad (66)$$

which ensured that $-d \ln(w_A^{\text{core}}(r_A))/dr_A \geq 2/\text{bohr}$, where Δr_A is the distance between adjacent radial shells. The core electron distributions of all isolated atoms naturally satisfy this constraint, because francium contains the most diffuse core electrons of all elements with a decay exponent of 2.14/bohr. This reduced the number of core fitting iterations to less than 30 (from an initial value of ~ 200 in some cases) with no significant effect on the resulting atomic charges.

Three radial shell and cutoff radius combinations are currently available in the program at ddec.sourceforge.net: (a) 50 radial shells with a 3 \AA cutoff, (b) 75 radial shells with a 4 \AA cutoff, and (c) 100 radial shells with a 5 \AA cutoff. Examining the c2 reference densities¹⁴ for different elements in the periodic table shows that 75 radial shells with a 4 \AA cutoff is a good choice for nearly all chemical systems. Specifically, the most diffuse neutral atoms with atomic numbers ≤ 109 are francium and radium, which have 0.11 electrons between 4 and 5 \AA . Commonly, Fr is in the +1 charge state, where it has only 0.0002 electrons between 4 and 5 \AA . The O^{2-} and Te^{2-} anions (which are among the most diffuse of the commonly occurring anions) have 0.003 and 0.028 electrons between 4 and 5 \AA , respectively. Ions like Cr^- and Mo^- are more diffuse with 0.048 and 0.053 electrons between 4 and 5

Å, respectively; however, these anions are not commonly occurring. Thus, a 4 Å cutoff is sufficient to converge the NACs and ASMs to within approximately ± 0.01 electrons for almost all chemical systems. A 5 Å cutoff is available for cases containing extremely diffuse atoms. For a given unit cell, the time required to compute NACs and ASMs scales linearly with the number of grid points per atom which in turn scales linearly with the cube of the cutoff radius. Consequently, using a 4 Å cutoff takes about twice as long as a 3 Å cutoff and about half as long as a 5 Å cutoff. Tests were performed for $\chi_{\text{spin}} = 1/2$ using TiI placed in the center of a $10 \times 10 \times 10$ Å super cell. For different cutoff values, the ratio of times to compute NACs was $\text{time}(3 \text{ Å})/\text{time}(4 \text{ Å}) = 0.47$ and $\text{time}(5 \text{ Å})/\text{time}(4 \text{ Å}) = 1.72$, and that to compute ASMs was $\text{time}(3 \text{ Å})/\text{time}(4 \text{ Å}) = 0.43$ and $\text{time}(5 \text{ Å})/\text{time}(4 \text{ Å}) = 1.77$. NACs differed by ± 0.01 (5 Å) and ± 0.05 (3 Å) relative to the 4 Å values. The iodine ASM changed by less than 0.002 when the cutoff radius was changed. The Ti ASM decreased from 3.076 (5 Å) to 3.066 (4 Å) to 3.001 (3 Å). The total spin moment was 2.930 (3 Å), 2.990 (4 Å), and 2.999 (5 Å) compared to the exact value of 3, which shows that about 0.07 and 0.01 polarized electrons were missed by the 3 and 4 Å cutoffs, respectively. For all of the other iodides in the Results below, <0.01 polarized electrons were missed by the 4 Å cutoff.

Bader atomic volumes were computed by the program of Henkelman and co-workers using the total electron density; then the valence and spin densities were integrated over these volumes to determine the atomic charges and ASMs, respectively.^{69–71} Bader's method can sometimes yield compartments that do not include an atomic nucleus (i.e., non-nuclear attractors).⁷² For the all-electron calculations studied in this paper, the number of Bader compartments was the same as the number of atoms, and there were no non-nuclear attractors. For calculations with RECPs (i.e., LANL2DZ and SDD basis sets), the core electron density replaced by the RECP was not included in the Bader analysis, and this resulted in non-nuclear maxima and several compartments per atom. We found that reliable Bader NACs and ASMs were obtained by assigning each compartment to the nearest nucleus, as done in the program of Henkelman and co-workers.^{69–71} In both Bader and DDEC analysis, the typical mesh size was ~ 0.05 Å, which provided <0.1 error in the total number of valence electrons. The valence electron density was then rescaled to provide the exact number of valence electrons. Natural population analysis was performed using the NBO 3.0 program⁷³ within Gaussian. Integration of the total spin density within an effective atomic sphere radius was performed using the Lorbit method in VASP using the default radii.

Attempts to compute Mulliken populations projected onto a minimal basis set (MBS) had limited success. As currently implemented in Gaussian 09, MBS population analysis projects onto a STO-3G basis set, which is only defined for atomic numbers ≤ 53 .^{15,64} Consequently, MBS analysis could not be performed for $[\text{GdI}]^{2+}$. The MBS method also requires the basis set used to compute the SCF wave function to be larger than the STO-3G basis set. This can become a critical problem when a RECP is used, since in this case the basis set used to compute the SCF wave function only needs to model valence electrons while the STO-3G basis set is all-electron. Consequently, MBS analysis could not be performed for MgI, MoI, SnI, or TeI when using SDD basis sets (which includes RECPs). It may be possible to fix some of these issues by treating valence and core electrons separately and projecting onto a different basis set than STO-3G. We were able to perform MBS analysis for TiI, the Cu_2

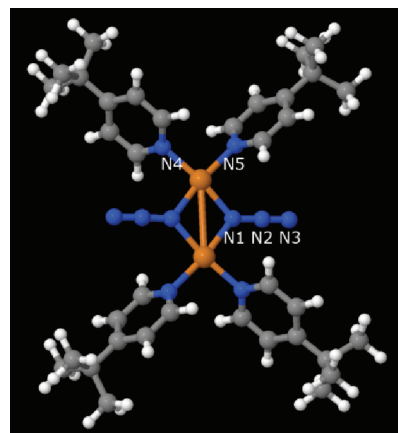


Figure 1. Cu_2 organometallic complex with triplet ground state and +2 net charge (atoms: Cu (gold), N (blue), C (gray), H(white)).

organometallic complex shown in Figure 1, and $[\text{Cr}(\text{CN})_6]^{3-}$. For TiI using SDD basis sets, the MBS NACs were 0.26 (Ti) and -0.26 (I), and the MBS ASMs were 3.14 (Ti) and -0.26 (I), which are reasonable. For the complex in Figure 1, the MBS NACs were 0.50 (Cu), -0.30 to 0.06 (C atoms), 0.09 to 0.16 (H atoms), and -0.33 to 0.15 (N atoms), and the MBS ASMs were 0.61 (Cu), -0.01 to 0.01 (C atoms), 0.00 (H atoms), and -0.03 to 0.13 (N atoms), which are reasonable. For $[\text{Cr}(\text{CN})_6]^{3-}$ with 6-311+G* basis sets, the MBS NACs were -11.97 (Cr), -2.45 (four C's), -1.96 (two C's), 4.12 (four N's), and 4.11 (two N's), and the MBS ASMs were 0.02 (Cr), 0.69 (four C's), 0.20 (two C's), -0.04 (four N's), and -0.02 (two N's), which are unreasonable. Because of these difficulties, we do not consider MBS population analysis further in this paper.

For systems with collinear magnetism, noncollinear DFT calculations and ASM analysis should yield the same results as collinear DFT calculations and ASM analysis. As an example, we compared noncollinear vs collinear DFT calculations and ASM analysis for a MgI molecule placed in the center of a $10 \text{ Å} \times 10 \text{ Å} \times 10 \text{ Å}$ super cell. Periodic DFT calculations were performed in VASP using the PBE XC functional.⁷⁴ The optimized collinear and noncollinear energies differed by less than 1 meV. Both collinear and noncollinear DFT calculations gave an optimized $M = 1.000$. For each of the $\chi_{\text{spin}} = 1.0, 3/14,$ and $1/2$, and LSF methods, the ASM magnitudes computed with noncollinear analysis differed by <0.001 from those computed using collinear analysis. These results confirm noncollinear and collinear analysis yield essentially identical results for a system with collinear magnetism.

4. RESULTS AND DISCUSSION

4.1. Diatomic Iodides. We first consider a series of diatomic iodides XI $\{X = \text{Mg}, \text{Ti}, \text{Mo}, \text{Sn}, \text{Te}, \text{Gd}^{2+}\}$ containing unpaired electrons. We specifically chose elements with electronegativity lower than iodine, so that one electron would be transferred to fill iodine's valence shell. Consequently, nearly all of the unpaired electrons reside on the X atom. This allows us to assess the accuracy of different ASM methods. Second, we computed the relative mean absolute error (RMAE) for fitting the magnetic field $\vec{B}^{\text{spin}}(\vec{r})$ in the volume between the surfaces defined by $3 \times$ and $4 \times$ vdW radii. Thus, the simplicity of these diatomic iodides allows us to objectively assess the accuracy of different ASM

Table 1. Atomic Spin Analysis of Diatomic Iodides

	total spin		atomic spin moment on cation												
			nonperiodic system ^a						periodic system ^b						
	other ^c	Lorbit ^d	Δ_S	Bader	NPA	χ_{spin}				Bader	Lorbit	χ_{spin}			
1.0						3/14	1/2	LSF	1.0			3/14	1/2	LSF	
MgI	1	0.23	0.00	0.85	0.87	0.85	0.99	0.93	1.04	0.85	0.14	0.85	1.02	0.94	1.06
TiI	3	2.26	0.02	3.03	3.13	3.01	3.25	3.14	3.29	2.96	2.29	2.93	3.17	3.07	3.22
MoI	5	4.00	0.03	4.95	5.00	4.87	5.12	5.01	5.17	4.89	4.00	4.82	5.09	4.97	5.13
SnI	1	0.54	0.00	0.94	1.00	0.94	1.03	0.99	1.06	0.94	0.52	0.92	1.02	0.98	1.05
TeI	1	0.52	0.00	0.85	0.87	0.84	0.88	0.86	0.89	0.81	0.41	0.80	0.84	0.82	0.86
[GdI] ²⁺	7	6.88	0.45	6.64	6.60	6.65	6.63	6.64	6.63	6.93	6.85	6.92	6.94	6.93	6.94
RMAE for the nonperiodic system [relative rank in brackets]															
zero ASM MAE (gauss) ^e		Bader	Lorbit	NPA	$\chi_{\text{spin}} = 1.0$	$\chi_{\text{spin}} = 3/14$	$\chi_{\text{spin}} = 1/2$	LSF							
MgI	32	0.249	0.813	0.237	0.249	0.174	0.201	0.161	[5]	[7]	[4]	[5]	[2]	[3]	[1]
TiI	88	0.050	0.257	0.038	0.053	0.034	0.037	0.041	[5]	[7]	[3]	[6]	[1]	[2]	[4]
MoI	144	0.027	0.204	0.021	0.038	0.010	0.020	0.014	[5]	[7]	[4]	[6]	[1]	[3]	[2]
SnI	22	0.085	0.472	0.060	0.085	0.059	0.063	0.068	[5]	[7]	[2]	[5]	[1]	[3]	[4]
TeI	21	0.043	0.491	0.038	0.047	0.037	0.040	0.039	[5]	[7]	[2]	[6]	[1]	[4]	[3]
[GdI] ²⁺	194	0.0038	0.0340	0.0052	0.0040	0.0039	0.0038	0.0039	[1]	[7]	[6]	[5]	[3]	[1]	[3]
avg.									[4.3]	[7.0]	[3.5]	[5.5]	[1.5]	[2.7]	[2.8]
st. dev.									[1.6]	[0.0]	[1.5]	[0.5]	[0.8]	[1.0]	[1.2]
mean abs difference between cation ASM and total spin moment															
		0.13	0.63	0.13	0.14	0.15	0.12	0.17							

^a Nonperiodic system using SDD basis sets. ^b Periodic system using $10 \text{ \AA} \times 10 \text{ \AA} \times 10 \text{ \AA}$ cubic unit cell and PAW method. ^c Total spin of the molecule and sum of ASMs for all methods except Lorbit. ^d Sum of ASMs for the Lorbit method. ^e 1 gauss = 10^{-4} tesla.

methods for simultaneously fitting $\vec{B}^{\text{spin}}(\vec{r})$ and atomic chemical states.

Table 1 shows results computed with the PBE functional. For the nonperiodic calculations, SDD basis sets were used that included RECPs for I (46 core), Ti (10 core), Mo (28 core), Sn (46 core), Te (46 core), and Gd (28 core). Comparison is made to periodic calculations using the PAW method with a $10 \text{ \AA} \times 10 \text{ \AA} \times 10 \text{ \AA}$ super cell. Since iodine is not very magnetic, the cation ASM should be within $\sim \pm 0.2$ of the total spin. The Bader; NPA; LSF; and $\chi_{\text{spin}} = 1, 3/14,$ and $1/2$ methods gave ASMs in this expected range for all materials in Table 1 except when SDD basis sets were used for [GdI]²⁺. Only for [GdI]²⁺ were the SDD and PAW results significantly different, with PAW giving the expected moment and SDD giving a moment ~ 0.35 different than expected. Consistent with these findings, Δ_S for SDD basis sets showed significant spin contamination ($\Delta_S = 0.45$) for [GdI]²⁺, and negligible spin contamination ($\Delta_S \leq 0.03$) for the other materials. ASMs computed by integrating the total spin density within an effective atomic sphere radius (denoted Lorbit) performed poorly because regions between the atomic spheres were not assigned to any atom, so the ASMs do not sum to the

total spin moment of the unit cell. For example, the sum of Lorbit ASMs for MgI was 0.23 compared to the correct value of 1. Table 1 also shows the relative rank of each method as defined by the computed RMAE values. The Lorbit method had a rank of 7 for each material, meaning it always gave the least accurate results among the seven ASM methods. Therefore, we do not consider Lorbit ASMs further in this paper.

Since decreasing the χ_{spin} value increases the proportion of spherical averaging in the optimization functional, smaller χ_{spin} values gave an improved prediction of $\vec{B}^{\text{spin}}(\vec{r})$, as evidenced by the smaller RMAE values. Among the seven ASM methods tested, using $\chi_{\text{spin}} = 3/14$ gave the most accurate prediction of $\vec{B}^{\text{spin}}(\vec{r})$. The $\chi_{\text{spin}} = 1/2$ and LSF methods were the next most accurate at predicting $\vec{B}^{\text{spin}}(\vec{r})$. Since RMAE is defined as the ratio of the mean absolute error (MAE) of the ASM model to the MAE when all ASMs were set to zero, the product of the RMAE and the zero ASM MAE recovers the computed MAE for each method. For example, the RMAE of MoI for the $\chi_{\text{spin}} = 1/2$ method was 0.020, and the zero ASM MAE was 144 gauss. The MAE computed from eq 24 was $0.020 \times 144 \text{ gauss} = 2.88 \text{ gauss}$ for MoI using $\chi_{\text{spin}} = 1/2$. In other words, ASMs computed using

Table 2. Atomic Charge Analysis of Diatomic Iodides

	optimized	net atomic charge on cation		
	d_0 (Å)	Bader	NPA	DDEC
MgI	2.69 ^a [2.61] ^b	0.68 ^a [0.76] ^b	0.61 ^a	0.41 ^a [0.42] ^b
TiI	2.61 ^a [2.61] ^b	0.51 ^a [0.58] ^b	0.40 ^a	0.32 ^a [0.37] ^b
MoI	2.67 ^a [2.62] ^b	0.63 ^a [0.60] ^b	0.39 ^a	0.32 ^a [0.34] ^b
SnI	2.85 ^a [2.74] ^b	0.26 ^a [0.29] ^b	0.39 ^a	0.26 ^a [0.23] ^b
TeI	2.80 ^a [2.69] ^b	0.11 ^a [0.02] ^b	0.06 ^a	0.02 ^a [0.00] ^b
[GdI] ²⁺	2.73 ^a [2.68] ^b	1.82 ^a [1.66] ^b	1.86 ^a	1.91 ^a [2.05] ^b
	RMSE no NACs (kcal/mol)	RRMSE for the nonperiodic system		
MgI	11.8	0.81	0.64	0.22
TiI	12.8	0.52	0.44	0.47
MoI	11.5	0.80	0.35	0.35
SnI	6.3	0.43	0.81	0.43
TeI	2.6	1.22	0.99	0.96
[GdI] ²⁺	41.5	0.13	0.15	0.17
avg.	14.4	0.65	0.56	0.43
st. dev.	13.8	0.38	0.31	0.28

^aNonperiodic system using SDD basis sets. ^bPeriodic system using $10 \text{ \AA} \times 10 \text{ \AA} \times 10 \text{ \AA}$ cubic unit cell and PAW method.

$\chi_{\text{spin}} = 1/2$ predicted $\bar{B}^{\text{spin}}(\vec{r})$ 50 times more accurately than when ASMs for all atoms were set to zero. The accuracy of the methods for reproducing chemical states can be estimated by considering iodine to be an essentially nonmagnetic atom, which means the cation ASM should be close to the total spin moment. Among the seven ASM methods, using $\chi_{\text{spin}} = 1/2$ gave the smallest mean absolute difference between the cation ASM and the total spin moment, suggesting that its accuracy is high. The Bader and NPA methods gave the next lowest values, suggesting they also have high accuracy. As shown in the rest of this paper, for most materials, the $\chi_{\text{spin}} = 1/2$, Bader, and (where applicable) NPA methods give similar ASMs. Overall, these results suggest that using $\chi_{\text{spin}} = 1/2$ gives an excellent combination of accuracy for reproducing both $\bar{B}^{\text{spin}}(\vec{r})$ and the chemical states of atoms in a material. This finding is encouraging, because the optimization functional in eq 29 is designed explicitly for this dual purpose.

Table 2 compares the Bader, NPA, and DDEC NACs for these same materials. When DDEC ASMs are computed using $\chi_{\text{spin}} = 1/2$ as discussed above, the variation in NACs between Bader, NPA, and DDEC methods is larger than the variation in ASMs between these methods. This agrees with previous findings that NACs are more sensitive to the choice of population analysis method than ASMs are.⁵⁰ Table 2 also lists the relative root mean squared error (RRMSE) in $V(\vec{r})$ for each of the point charge models. RRMSE is defined as the root mean squared error (RMSE) of $V(\vec{r})$ for a point charge model divided by the RMSE when all NACs are set to zero, where grid points for computing RMSE are uniformly distributed between the surfaces defined by $1.4\times$ and $2.0\times$ vdW radii.^{16,54,75} The RRMSE values in Table 2 suggest that the DDEC method is the most accurate for reproducing $V(\vec{r})$, which agrees with our previous findings.¹⁴

4.2. Comparison to Experimental Atomic Spin Moments.

We now compare theoretically computed ASMs to experimental data for several molecules and crystals. Table 3 compares experimental^{76–78} and computed ASMs for the Cu_2 organometallic complex shown in Figure 1, $[\text{Cr}(\text{CN})_6]^{3-}$, and solid Fe_3Si .

Geometry optimization was performed for both the singlet and triplet states of the Cu_2 complex shown in Figure 1. The minimum energy singlet state was a closed shell singlet (all ASMs were zero) with an energy 0.78 eV above the minimum energy triplet state. This agrees with experiments showing that the ground state is a triplet.⁷⁶ Solid Fe_3Si has adjacent magnetic atoms with overlapping electron distributions. As mentioned above, the LSF method performs poorly for this type of material; specifically, LSF predicts the wrong sign for the ASM of Si and takes 2094 iterations to converge. For all three materials, the Bader and DDEC ASMs are close to each other and in good agreement with experiments. The value of χ_{spin} had a small effect on the computed ASMs but a large effect on the number of required iterations. For the two molecular systems, the small amount of spin contamination, $\Delta_S \leq 0.03$, shows that DFT accurately reproduced the spin states. For these two systems, the NPA ASMs were similar to those of the other population analysis methods. NACs for all three materials are listed in the Supporting Information. For $[\text{Cr}(\text{CN})_6]^{3-}$, the Cr NAC was 1.45 (Bader), 0.28 (DDEC), and -0.78 (NPA), which clearly shows that NACs are more sensitive than ASMs to the choice of population analysis method. The RMSE values for $[\text{Cr}(\text{CN})_6]^{3-}$ were 1.3 (DDEC), 2.2 (NPA), 9.0 (Bader), and 16.9 (no charges) kcal/mol, which shows that DDEC was the most accurate for reproducing $V(\vec{r})$.

We now consider the computation of ASMs for a highly correlated material. Magnetite, which has an inverse spinel structure, has one tetrahedrally coordinated Fe atom and two octahedrally coordinated Fe atoms per formula unit.⁷⁹ The Fe(tet) atom has a formal oxidation state of +3. There are two possibilities for the Fe(oct). The first possibility is for the two Fe(oct) to be equivalent with a formal oxidation state of +2.5, and the second possibility is for one Fe(oct) to have a +2 oxidation state with the other having a +3 oxidation state.⁷⁹ Bulk magnetite undergoes a structural and electronic transition near $T_V = 120 \text{ K}$ called the Verwey transition,⁸⁰ whose precise mechanism is still debated.^{81,82} Above T_V , electrons move rapidly between Fe(oct)s, causing them to appear equivalent with an effective +2.5 oxidation state. Progress has been made to use the ferrimagnetic half-metallic state of bulk magnetite above T_V for spintronics applications.^{83–86} The redox properties of magnetite surfaces lead to unusual adsorption properties,^{87,88} and a Verwey-like charge-spin ordering transition accompanies the magnetite (001) surface reconstruction.^{89,90} Below T_V , the Fe(oct)s, in bulk magnetite become inequivalent and the crystal structure changes. As previously reported, the Hubbard on-site parameter U_{eff} offers a convenient way to artificially transition between these two regimes, because in a DFT +U calculation the Fe(oct) sites appear equivalent for $U_{\text{eff}} < 2.6 \text{ eV}$ but inequivalent for $U_{\text{eff}} > 2.6 \text{ eV}$.⁷⁹ Table 4 shows the effect of the Hubbard on-site parameter on bulk magnetite's charge and spin ordering, using the PBE functional⁷⁴ and Dudarev et al.'s⁹¹ DFT +U method. The lattice constant was optimized for each U_{eff} value using a cubic Fe_6O_8 unit cell, which is the experimental unit cell shape above T_V . As apparent from the NACs and ASMs shown in Table 4, the two Fe(oct) sites were equivalent for $U_{\text{eff}} = 0$ and inequivalent for $U_{\text{eff}} = 3.2 \text{ eV}$. For $U_{\text{eff}} = 3.2 \text{ eV}$, this also causes the O atoms to segregate into two groups with slightly different Fe–O distances (not shown), NACs, and ASMs. For $U_{\text{eff}} = 3.2 \text{ eV}$, the $\text{Fe}^{\text{III}}(\text{tet})$ and $\text{Fe}^{\text{III}}(\text{oct})$ had the same Bader (~ 1.6) or DDEC (~ 2.0) NACs, which rationalizes an assignment of +3 oxidation for both sites. The experimental value for the Fe(tet) ASM is -3.82 .⁹² Irrespective of the population analysis method, our computations gave Fe(tet) ASM values slightly lower than this

Table 3. Comparison of Calculated and Experimental Atomic Spin Moments

	exp. value	NPA	Bader	$\chi_{\text{spin}} = 1.0$	$\chi_{\text{spin}} = 3/14$	$\chi_{\text{spin}} = 1/2$	LSF
Fe ₃ Si solid (PBE functional, optimized spin moment per unit cell = 5.09)							
Fe	2.20 ^b	n.a.	2.60	2.50	2.54	2.52	2.25
Fe	1.35 ^b	n.a.	1.31	1.31	1.31	1.31	1.27
Fe	1.35 ^b	n.a.	1.31	1.31	1.31	1.31	1.27
Si	-0.07 ^b	n.a.	-0.14	-0.04	-0.07	-0.06	0.30
iterations ^a				1	17	7	2094
[Cr(CN) ₆] ³⁻ (B3LYP/6-311+G*, total spin moment = 3, $\Delta_S = 0.03$)							
Cr	3.25 ^c	2.96	2.85	2.81	2.97	2.90	3.10
C	-0.09 ^c	-0.06	-0.03	-0.02	-0.06	-0.04	-0.10
N	0.05 ^c	0.07	0.06	0.05	0.06	0.06	0.08
iterations ^a				1	17	9	66
Cu ₂ organometallic complex (B3LYP/LANL2DZ, total spin moment = 2, $\Delta_S = 0.01$)							
Cu	0.77 ^d	0.53	0.59	0.59	0.58	0.59	0.56
N1	0.07 ^d	0.14	0.11	0.11	0.12	0.11	0.13
N2	-0.02 ^d	-0.04	-0.02	-0.01	-0.03	-0.02	-0.04
N3	0.05 ^d	0.13	0.12	0.11	0.13	0.12	0.13
N4	0.07 ^d	0.12	0.10	0.09	0.10	0.09	0.12
N5	0.05 ^d	0.12	0.10	0.09	0.10	0.09	0.12
other atoms	$\leq 0.03^d$	≤ 0.01	<0.01	<0.01	<0.01	<0.01	<0.02
iterations ^a				1	12	7	64

^a Iterations required to converge ASMs to within 5×10^{-5} . ^b From ref 78. ^c From ref 77. ^d From ref 76.

Table 4. Effect of XC Functional on Bulk Magnetite's Charge and Spin Ordering

	oxidation state	number per unit cell	NAC		ASM			LSF	
			Bader	DDEC	Bader	$\chi_{\text{spin}} = 1.0$	$\chi_{\text{spin}} = 3/14$		$\chi_{\text{spin}} = 1/2$
$U_{\text{eff}} = 0$ (optimized lattice constant = 8.303 Å)									
Fe(oct)	2.5	4	1.46	1.53	3.56	3.43	3.52	3.49	3.33
Fe(tet)	3	2	1.46	1.74	-3.48	-3.32	-3.49	-3.42	-3.66
O	-2	8	-1.09	-1.20	0.09	0.12	0.11	0.11	0.25
iterations ^a						1	17	8	1130
$U_{\text{eff}} = 3.2$ eV (optimized lattice constant = 8.435 Å)									
Fe(oct)	3	2	1.62	1.97	4.06	3.90	4.01	3.96	3.96
Fe(oct)	2	2	1.40	1.47	3.65	3.52	3.61	3.57	3.60
Fe(tet)	3	2	1.59	1.99	-3.98	-3.81	-4.00	-3.91	-4.10
O	-2	4	-1.15	-1.38	0.10	0.13	0.12	0.12	0.17
O	-2	4	-1.16	-1.34	0.04	0.07	0.06	0.06	0.09
iterations ^a						1	17	8	322
HSE06 (optimized lattice constant = 8.35 Å)									
Fe(oct)	>2.5	2	1.66	1.89	3.96	3.82	3.93	3.88	3.87
Fe(oct)	<2.5	2	1.63	1.84	3.91	3.77	3.87	3.83	3.85
Fe(tet)	3	2	1.68	2.11	-4.02	-3.84	-4.02	-3.94	-4.12
O	-2	4	-1.23	-1.45	0.02	0.04	0.04	0.04	0.09
O	-2	4	-1.25	-1.47	0.06	0.08	0.07	0.08	0.11
iterations ^a						1	15	8	555

^a Iterations required to converge ASMs to within 5×10^{-5} .

for $U_{\text{eff}} = 0$. On the other hand, for $U_{\text{eff}} = 3.2$ eV, the computed Fe(tet) ASM was equal to or slightly greater than the experimental value, which suggests that $0 < U_{\text{eff}} < 3.2$ eV is probably optimal for the description of charge and spin ordering in bulk

magnetite. For both $U_{\text{eff}} = 0$ and 3.2 eV, the total spin moment per Fe₃O₄ formula unit was +4, and recent experiments gave approximately the same value.⁹³ The main conclusion from these results is that for highly correlated materials the XC theory has a

larger effect on the computed ASMs than the population analysis method does.

These results reflect the tendency of common GGA functionals to overly delocalize electrons in many extended materials.^{94,95} Specifically, the exact XC functional should vary linearly in the total number of electrons when the total number of electrons is fractional.⁹⁶ XC functionals giving convex behavior overly delocalize electrons in extended systems, while those giving concave behavior overly localize electrons in extended systems.⁹⁷ DFT+U is one method for approximately correcting delocalization errors.⁹⁸ Because $U_{\text{eff}} \gg 0$ favors the localization of electrons,⁹⁹ combining a functional like PBE or PW91 with a suitably chosen U_{eff} value can approximately cancel the delocalization error. The U_{eff} value can be different for different oxidation states and environments of the same element.⁹⁹ In the DFT+U example above, $U_{\text{eff}} = 3.2$ eV applies only to the Fe atoms, and $U_{\text{eff}} = 0$ for the oxygen atoms. The Hartree–Fock method has concave behavior, leading to overly localized electrons.⁹⁷ Thus, using a fraction of Hartree–Fock exchange to create a hybrid functional is another way to partially correct the delocalization error.^{95,98,100} For range-separated hybrids, the fraction of HF exchange varies as a function of interelectron distance. The range-separated hybrid HSE06 improves performance in extended systems by mixing 75% PBE with 25% HF exchange at short interelectron distances and by using PBE alone at large interelectron distances.¹⁰¹ For extended systems with conducting electrons, this has a lower computational cost and improved convergence with respect to k -point sampling than using the same fraction of HF exchange at all interelectron distances.^{102,103} The NACs and ASMs for magnetite computed with HSE06 are shown in Table 4. Due to the higher computational cost associated with including HF exchange, this HSE06 calculation was performed only at the Γ point. As evident from the higher magnitude Fe ASMs in Table 4, the spin density is more localized for HSE06 than for PBE. The magnitude of the NACs is also higher for HSE06 than for PBE. The amount of charge and spin disproportionation between the different types of Fe(oct) sites is higher for PBE+U than for HSE06. For HSE06, the amount of charge disproportionation was <0.1 electron, while for PBE+U it was several tenths of an electron. The magnitude of the oxygen NAC was higher for HSE06 than for PBE+U. In summary, these results show that DFT+U and HSE06 increased the localization of charge and spin compared to PBE for the Fe atoms in magnetite.

4.3. A Single Molecule Magnet with Highly Noncollinear Magnetism. Noncollinear magnetism occurs when the spin magnetization density $\vec{m}(\vec{r})$ is not parallel to a global axis. Previous studies have shown the spin magnetization direction changes significantly only where the spin magnetization magnitude is small. Specifically, the spin magnetization direction stays approximately constant inside a magnetic atom but may change direction near the boundary of a magnetic atom or inside nonmagnetic atoms.^{22,57–59} Since there is little variability in the magnetization direction inside individual magnetic atoms, the ASMs describe most of the essential information about the magnetic structure. As an example of noncollinear magnetism, we study the ferrous cube complex $\text{Fe}_4\text{C}_{40}\text{H}_{52}\text{N}_4\text{O}_{12}$ (commonly called $[\text{Fe}_4(\text{sae})_4(\text{MeOH})_4]$) synthesized and experimentally characterized by Oshio et al., which was the first example of a single-molecule magnet involving an Fe(II) cluster.^{104,105} This material contains Fe and O atoms on alternating corners of a distorted cube surrounded by organic ligands. Experiments showed that this material is a single molecule magnet with an activation

energy, called the magnetic anisotropy barrier, of 2.4 meV (which corresponds to 28.4 K) for reorientation of the magnetization.^{104,105} This was interpreted as a ground state in which the spin moments of two Fe atoms are antiparallel along a first axis and the spin moments of the remaining two Fe atoms are antiparallel along a second axis approximately perpendicular to the first.¹⁰⁴ To better quantify the magnetic structure of this complex, we performed noncollinear DFT calculations in VASP using the PW91 functional with the correlation interpolation of Vosko et al.^{57,61,106,107} For these simulations, the molecule was placed in the center of a $19 \times 19 \times 19 \text{ \AA}^3$ super cell with calculations performed at the Γ point. (See the Supporting Information for additional details.)

Finding the ground state of systems with highly noncollinear magnetism is challenging because the ASMs may point in any direction. Noncollinear calculations are commonly performed using the converged collinear magnetic structure as an initial guess.¹¹ This is only reliable, however, if the amount of noncollinearity is small. We developed the procedure shown in Figure 2 for finding the ground state of systems with highly noncollinear magnetism. In this procedure, spin–orbit coupling is turned off until the final geometry and four electron density components have been determined. First, the spin moments of magnetic atoms are initially set to random values, while those for nonmagnetic atoms are initially set to zero. In this example, the x , y , and z components of each Fe atom's spin moment were initialized to random values in the continuous interval between -4 and $+4$, while the spin moments for the remaining atoms were initialized to zero. The next step is to compute the energy using several such sets of random spin moments on the initial geometry. Configurations with comparatively low energy are then selected for subsequent geometry optimization. In this example, we used 12 initial sets of random spin moments and selected the six lowest energy ones for subsequent geometry optimization. Convergence of noncollinear magnetic structures can be slow because the energies for relative rotations of the spin moments can be small. As a result, the converged geometry with the lowest energy was selected for further optimization of the magnetic degrees of freedom. In this example, we continued to optimize the magnetic degrees of freedom by performing 10 ionic steps in which the atomic positions were held constant with 200 SCF cycles per ionic step, for a total of 2000 SCF cycles at the converged geometry. During this process, the molecule's total spin moment magnitude relaxed from 0.48 to 0.09. After relaxation of the magnetic degrees of freedom, the forces were rechecked, and the geometry was rereaxed if the forces were no longer converged. In this example, the forces were still converged to better than 0.03 eV/\AA , so re-relaxation of the geometry was unnecessary. This process yields the final converged geometry, four charge density components, and wave function, which are stored for use in the spin–orbit coupling calculations. Because the spin–orbit coupling energy is small, the spin–orbit coupling calculations used this saved wave function as an initial estimate and kept the geometry and four charge density components constant, except for a rotation of the spin axes. The set of possible rotations of the spin axes corresponds to the set of unit vectors from the center of a unit sphere to its surface. The spin–orbit coupling energy was computed for 60 uniformly spaced rotations defined by a grid of 60 points on the unit sphere comprising the 60 vertices of a truncated icosahedron.

After determining the system's ground state, the values $\Xi^{\text{tot}} = 5.9871$ and $\Xi^{\text{mag}} = 5.9844$ au were calculated according to eqs S5 and S3 using the finite difference approximation applied to the

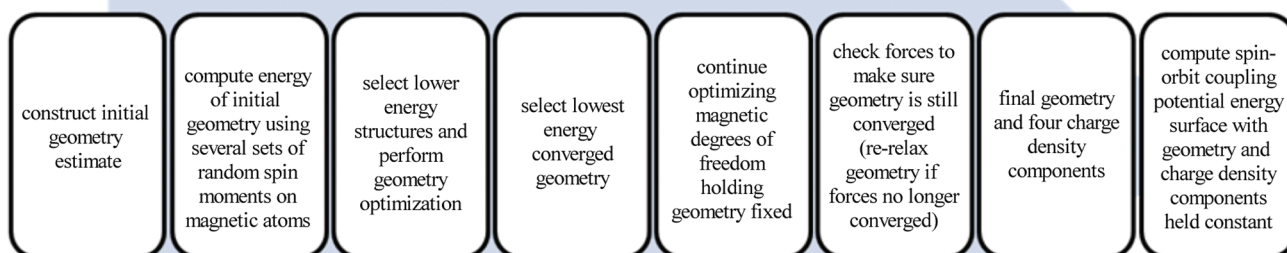


Figure 2. Method for computing the ground state of systems with highly noncollinear magnetism.

same grid of $\vec{m}(\vec{r})$ values used to compute the ASMs. This gave an estimated $\Xi^{\text{dir}} = 0.0027$, which is 3 orders of magnitude smaller than Ξ^{mag} . Almost all of Ξ^{tot} is due to $m(\vec{r})$ changes, with very little due to $\hat{m}(\vec{r})$ changes, because the magnetic Fe atoms in the complex are separated by nonmagnetic atoms leading to $\hat{m}(\vec{r})$ changes only in the space between Fe atoms where $m(\vec{r}) \approx 0$. Since Ξ^{dir} is negligible, the XC energy should be accurately represented by $E_{\text{XC}}^{\text{two}}$. Accordingly, a GGA functional like PW91 is appropriate for modeling noncollinear magnetism in this system.

Figure 3 shows the computed ground state magnetic structure and the spin–orbit coupling potential energy surface formed by globally rotating $\vec{m}(\vec{r})$. These spin–orbit coupling interactions determine the preferred orientation of the magnetic structure relative to the molecular structure in the absence of an externally applied magnetic field. Interestingly, Figure 3 contains three minima connected by three transition states. The computed activation barrier for converting between the three minima is 2.9 meV, in good agreement with the experimental magnetic anisotropy barrier of 2.4 meV. To the best of our knowledge, experiments have not determined the number of minima on the spin–orbit coupling potential energy surface.¹⁰⁴ For single molecule magnets with collinear magnetism, one expects an even number of energy minima due to the symmetry between up and down spins, and model Hamiltonians for single molecule magnets usually assume an energy well with two minima.¹⁰⁸ If confirmed, this example of a triple-minima energy well would be a significant finding. To the best of our knowledge, spin–orbit coupling potential energy surfaces with similar features have not been previously reported.

For each atom in the complex, Bader and DDEC analysis gave essentially identical ASMs; for each atom in the molecule $|M_A^{\text{Bader}} - M_A^{\text{DDEC}}| \leq 0.005$. Using $\chi_{\text{spin}} = 1/2$, the ASM magnitude was 2.06 for each Fe atom and ≤ 0.10 for all other atoms. (For $\chi_{\text{spin}} = 1.0$ (3/14), these values were 2.01 (2.10) and ≤ 0.09 (0.12), respectively.) Thus, there are approximately eight magnetic electrons. ASMs converged in 1, 8, and 15 iterations for $\chi_{\text{spin}} = 1, 1/2$, and 3/14, respectively. The computed magnitude of the molecule's total spin moment was ~ 0.1 . To the best of our

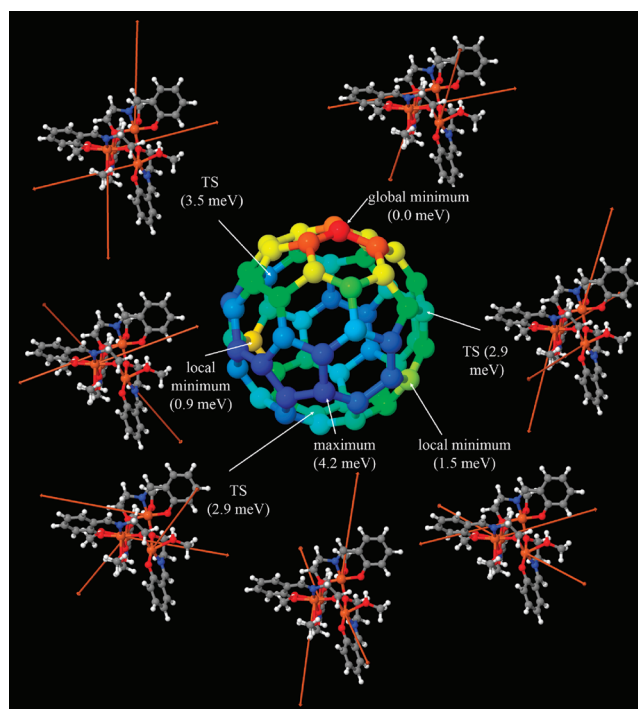


Figure 3. Spin–orbit coupling potential energy surface of the ferrous cube complex $\text{Fe}_4\text{C}_{40}\text{H}_{52}\text{N}_4\text{O}_{12}$. The center shows 60 points forming the truncated icosahedron used to sample the unit sphere of possible rotations of the spin axes, where color indicates the relative energy of each rotation. Adjacent to each labeled minimum, transition state, and maximum, the ASMs are displayed as vectors on the molecular structure, where color indicates the element (orange, Fe; red, O; blue, N; gray, C; white, H). ASMs are insignificant on all atoms except Fe. The large orange vectors show the direction and magnitude of Fe ASMs. The magnitudes and relative angles between ASMs are constant, but they rotate with respect to the molecular structure.

knowledge, experiments have not yet determined the molecule's total net spin moment at $\vec{B}_{\text{ext}} = 0$.¹⁰⁴ The antiparallel Fe ASMs

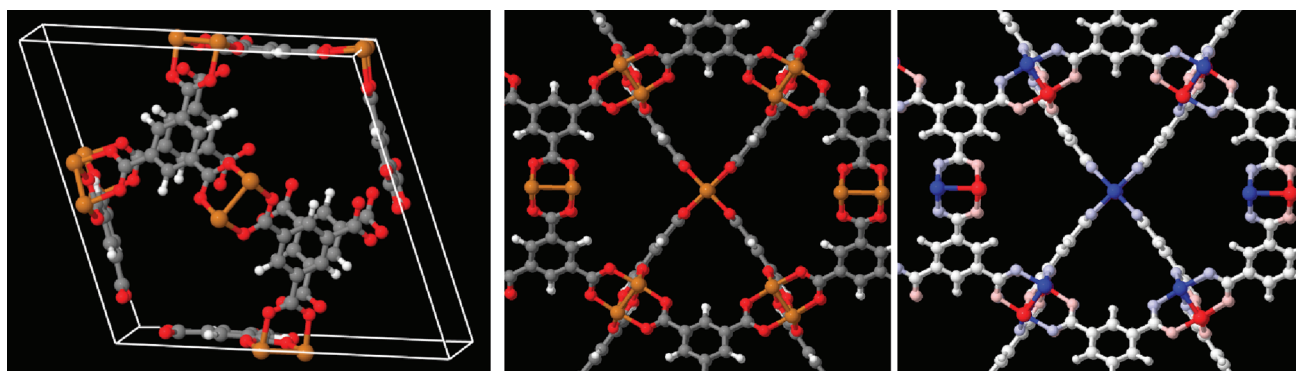


Figure 4. The metal organic framework $\text{Cu}_3(\text{BTC})_2$. Left: the unit cell (atoms: Cu (gold), O (red), C (gray), H(white)). Middle: looking down the pores. Right: atoms colored by ASM (-0.5 (blue) to 0.0 (white) to $+0.5$ (red)).

have computed angles of 178° ($\chi_{\text{spin}} = 1/2$), indicating almost perfect antiparallelism. The first and second sets of antiparallel Fe ASMs cross at a computed 10° deviation from perpendicular. This slight deviation from perpendicular is probably due to the distorted cubic geometry.

Our calculated noncollinear magnetic structure has some important similarities and differences to previous interpretations. Oshio et al. used simple crystal field theory and the angular overlap method to interpret the experimental results. Specifically, they assumed that each Fe(II) atom was occupied by six d electrons in a high spin arrangement having four unpaired electrons.^{104,109} They predicted 16 unpaired electrons ($S = 8$) for the entire molecule with a noncollinear ground state arrangement having two approximately perpendicular antiparallel pairs of Fe ASMs.¹⁰⁴ Our calculations show there are two approximately perpendicular antiparallel pairs of Fe ASMs in the ground state (See Figure 3) with approximately eight unpaired electrons in the molecule. Specifically, spherical harmonic projection in VASP gave 6.25 d electrons per Fe atom, and the computed ASMs showed only two d electrons per Fe atom are unpaired. Therefore, we believe the previously reported crystal field theory interpretation is incorrect, and there are eight rather than 16 unpaired electrons in the molecule. To the best of our knowledge, the only previous ab initio calculations performed for this complex are the collinear DFT calculations of Ribas-Arino et al.¹¹⁰ Ribas-Arino et al. did not perform calculations to determine the ground spin state but instead restricted calculations to the $S = 8$ state suggested by Oshio et al. Using this restriction, they computed an ASM magnitude of 3.6 for each Fe atom.¹¹⁰ In agreement with experiments, their results indicated that $[\text{Fe}_4(\text{sae})_4(\text{MeOH})_4]$ is a single molecule magnet, while the related iron cube system $[\text{Fe}_4(\text{sap})_4(\text{MeOH})_4]$ is not.^{104,110} They computed a magnetic anisotropy barrier of 0.5 meV for $[\text{Fe}_4(\text{sae})_4(\text{MeOH})_4]$, which is not nearly as close to the experimental value of 2.4 meV as our computed value of 2.9 meV. Our ab initio calculations are the first to consider noncollinear magnetism for this complex.

Experiments performed from near absolute zero temperature to above room temperature showed that $[\text{Fe}_4(\text{sae})_4(\text{MeOH})_4]$ has a positive magnetic susceptibility, meaning that the net magnetic moment increases in the direction of an applied magnetic field.^{104,105} Within the framework of the $S = 8$ spin model described above, Oshio et al. and Ribas-Arino et al. interpreted the positive magnetic susceptibility as an increasing occupation of the $S_z = +8$ state at the expense of the $S_z = -8$ state as the external magnetic field is increased.^{104,105,110}

Our noncollinear DFT computations provide a different explanation for the positive magnetic susceptibility. Specifically, our calculations showed that the net spin moment magnitude M could be changed from 0.09 to 0.48 with only a 0.17 meV increase in total energy, which is an order of magnitude lower than the magnetization reorientation energy barrier. The magnitude of each Fe ASM remained constant at 2.06 during this increase in M , while the angles between them changed slightly. This means that the Fe ASMs can undergo small relative rotations with little change in the total energy. These calculations were performed with no external magnetic field applied. When an external magnetic field \vec{B}_{ext} is applied, an energy of $g\mu_B M \cdot \vec{B}_{\text{ext}}/2$ must be included in the system's Hamiltonian; therefore, to minimize the system's energy in the presence of nonzero \vec{B}_{ext} , the Fe spin moments will rotate so as to increase $M/2$ in the direction of the applied magnetic field. Our results strongly suggest that the individual Fe ASMs in the noncollinear arrangement can easily rotate toward the direction of an externally applied magnetic field, leading to a change in the angles between Fe ASMs. In summary, our calculations for $\vec{B}_{\text{ext}} = 0$ give a ground state having (a) eight unpaired electrons with an ASM of ~ 2.06 on each Fe atom, (b) small spin moments on all atoms except Fe, (c) noncollinear magnetism in which four Fe spin moments are aligned in two pairs with antiparallel alignment within a pair and a deviation of $\sim 10^\circ$ from perpendicular between pairs, (d) a magnetization reorientation barrier of ~ 2.9 meV with a triple-minima energy, and (e) $M \approx 0.1$. For $\vec{B}_{\text{ext}} \neq 0$, a positive magnetic susceptibility is predicted due to canting of the noncollinear Fe spin moments. The computational techniques described here will also be useful to better understand the properties of other noncollinear magnetic systems.

4.4. The Antiferromagnetic Metal Organic Framework $\text{Cu}_3(\text{BTC})_2$. In the above sections, ASMs have been computed for nonporous solids and molecular systems. We now consider a porous solid. The metal–organic framework $\text{Cu}_3(\text{benzenetricarboxylate})_2$, abbreviated $\text{Cu}_3(\text{BTC})_2$, has open metal sites that can bind adsorbates and is widely studied for gas adsorption and catalytic applications.^{65,111–116} This material has high adsorption capacities for acetylene and nitric oxide,^{115,116} is a Lewis acid catalyst for several molecular rearrangement reactions,¹¹¹ and activates molecules for liquid phase cyanosilylation.¹¹⁴ Figure 4 shows the unit cell of $\text{Cu}_3(\text{BTC})_2$ containing 156 atoms. The copper atoms are arranged in pairs with a Cu–Cu bond length of ~ 2.5 Å. The distance between adjacent Cu pairs is ~ 10 Å. Oxygen atoms connect the Cu pairs to aromatic linking groups. Temperature-dependent magnetic susceptibility and electron

Table 5. Magnitude of ASMs for $\text{Cu}_3(\text{BTC})_2$

	oxidation state	NAC		ASM				
		Bader	DDEC	Bader	$\chi_{\text{spin}} = 1.0$	$\chi_{\text{spin}} = 3/14$	$\chi_{\text{spin}} = 1/2$	LSF
Cu	2	1.04	0.89	0.49	0.49	0.48	0.49	0.47
O	-2	-1.09	-0.56	0.09	0.09	0.09	0.09	0.10
other atoms				<0.01	<0.01	<0.01	<0.01	<0.01
iterations ^a					1	12	6	39

^a Iterations required to converge ASMs to within 5×10^{-5} .

Table 6. Population Analysis for Spin States of Ozone

method	total		relative	O–O–O	O–O	net	O– <u>O</u> –O	O– <u>O</u> –O	<u>O</u> –O–O	<u>O</u> –O–O
	S_z	Δ_S	energy (eV)	angle (deg)	dist. (Å)		charge	NAC ^a	ASM ^a	NAC ^b
singlet state										
exp.	0		0	116.8 ^c	1.27 ^c					
B3LYP	0	0	0	118.5	1.26	0	0.39	0.00	-0.20	0.00
PW91	0	0	0	118.4	1.28	0	0.37	0.00	-0.18	0.00
CCSD	0	0	0	117.8	1.24	0	0.37	0.00	-0.18	0.00
SAC–CI	0	0	0	117.8	1.24	0	0.40	0.00	-0.20	0.00
triplet states										
exp.	1		1.18 ^d	98.9 ^e	1.35 ^e					
B3LYP	1	0.01	0.82	99.2	1.32	0	0.20	0.38	-0.10	0.81
PW91	1	0.01	0.96	99.3	1.35	0	0.20	0.38	-0.10	0.81
CCSD	1	<0.08 ^g	1.06	97.5	1.32	0	0.23	0.42	-0.12	0.79
SAC–CI	0	0	1.30	97.4	1.32	0	0.20	0.00	-0.10	0.00
+1 cation doublet state										
exp.	1/2		12.52 ^f	131.5 ^f	1.25 ^f					
B3LYP	1/2	0.01	12.74	134.6	1.21	1	0.53	0.22	0.24	0.39
PW91	1/2	0.00	12.35	134.3	1.23	1	0.48	0.20	0.26	0.40
CCSD	1/2	<0.03 ^g	12.71	133.4	1.18	1	0.49	0.24	0.26	0.38

^a DDEC NAC and ASM ($\chi_{\text{spin}} = 1/2$) of the center oxygen atom. ^b DDEC NAC and ASM ($\chi_{\text{spin}} = 1/2$) of each outer oxygen atom. ^c Ref 121. ^d Ref 122. ^e Ref 123. ^f Ref 124. ^g Δ_S values for the HF reference configuration were used as an upper bound for the CCSD Δ_S values.

spin resonance experiments show that the Cu atoms have strong antiferromagnetic coupling within a pair and weak ferromagnetic coupling between pairs.^{117–119} Electron paramagnetic resonance (EPR) showed that the arrangement of ASMs was not static; there was dynamic spin exchange between the Cu atoms, as evidenced by the broad EPR peaks.¹¹⁹ Thus, a particular Cu atom is not locked into a spin-up or -down orientation but rather switches between the two on some short time scale. Watanabe and Sholl showed that the PW91 functional correctly predicts the antiferromagnetic ground state of this material.⁶⁵ Using this functional, the ASMs for Cu and O have magnitudes of 0.49 and 0.09, respectively, as shown in Table 5. The ASMs were not sensitive to the choice of population analysis method, but the NACs were. The sign of each ASM is shown on the right side of Figure 4.

4.5. Ozone. As a final example, we computed ASMs for singlet, triplet, and cation doublet states of ozone using DFT and coupled cluster expansions with 6-311+G* basis sets. For the coupled cluster expansions, CCSD was used to compute the singlet, $S_z = +1/2$ doublet, and $S_z = +1$ triplet states, while the symmetry adapted cluster configuration interaction (SAC–CI) method of Nakatsuji et al.^{45,46,120} was used to compute the singlet and $S_z = 0$

triplet states. SAC–CI is a coupled cluster expansion that includes symmetry operators for computing the $S_z = 0$ ground and excited states.^{45,46,120} Except for a flip of the spin direction, the $S_z = -1/2$ doublet and $S_z = -1$ triplet states (not shown) are identical to the $S_z = +1/2$ doublet and $S_z = +1$ triplet states, respectively. As shown in Table 6, for each spin state, the optimized geometries and relative energies are in good agreement with the experimental value. We find that the computed ASMs are similar for the CCSD and DFT methods. The small DFT Δ_S values indicate that the spin densities can be accurately reproduced by DFT. Our results suggest that DFT and CCSD give similar ASMs when Δ_S is small.

5. CONCLUSIONS

A functional for optimizing atomic spin moments (ASMs) using the DDEC atomic density distributions was presented that has a unique minimum. These ASMs are especially suitable for constructing interaction potentials for atomistic simulations, because they accurately reproduce the spin-derived magnetic field component, $\vec{B}^{\text{spin}}(\vec{r})$, outside a material's electron distribution as well as the chemical states of atoms in a material. The parameter

χ_{spin} , which determines the relative weights of proportional and spherically averaged atomic spin density in this functional, had a small influence on the ASM values. We recommend the value $\chi_{\text{spin}} = 1/2$, which gives converged ASMs in fewer than 10 iterations and is accurate for reproducing both $\vec{B}^{\text{spin}}(\vec{r})$ and the chemical states of atoms. This functional performed better than a functional that minimizes the least squared difference between the local and spherically averaged atomic spin magnetization densities.

Results showed that the computational method (i.e., DFT, DFT+U, coupled-cluster, etc.) has a larger influence on the computed ASMs than the population analysis method (i.e., Bader, NPA, DDEC). The Bader, DDEC, and (where applicable) NPA methods gave similar ASMs, but different net atomic charges. Bader and DDEC ASMs have the advantage of being applicable to periodic and nonperiodic materials and to collinear and noncollinear magnetism. The computed ASMs were in good agreement with available experimental data for a variety of periodic and nonperiodic materials. A range of materials exhibiting collinear magnetism were studied, including diatomic iodides, an antiferromagnetic metal organic framework, $\text{Cu}_3(\text{BTC})_2$, several ozone spin states, mono- and binuclear transition metal complexes, and ferri- and ferro-magnetic solids (Fe_3O_4 and Fe_3Si). The highly correlated material magnetite was studied using DFT+U and HSE06 methods. In agreement with prior literature, we found that the U_{eff} parameter has a large effect on charge and spin ordering in mixed oxidation state materials like magnetite. Both DFT+U and HSE06 separate octahedral Fe sites into nominally Fe^{II} and Fe^{III} atoms with different net charges and ASMs, while PBE calculations give only one type of octahedral Fe.

Noncollinear magnetic calculations are usually started using the converged collinear magnetic structure as an initial guess. This procedure is only reliable, however, when the amount of noncollinearity is small. We developed a procedure for finding the ground state of systems with highly noncollinear magnetism. The single-molecule magnet $\text{Fe}_4\text{C}_{40}\text{H}_{52}\text{N}_4\text{O}_{12}$ was studied by this method. The computed ground state magnetic structure was highly noncollinear and provided an improved explanation of the experimental results. Also, we briefly discussed the theory of XC functionals for studying noncollinear magnetism and introduced two descriptors, Ξ^{dir} and Ξ^{mag} , that are helpful for choosing an appropriate XC functional when studying noncollinear magnetism.

APPENDIX

1. Proof H Has a Unique Minimum. The core of the method we have introduced is the minimization of the functional H defined in eq 29. We now show that H has only one minimum. To prove uniqueness, it is sufficient to show that H is everywhere convex, which is necessarily true if its second-order variational derivative

$$\delta^2 H = \sum_A \sum_B \oint \oint \delta \vec{m}_A(\vec{r}_A) \cdot \frac{\partial^2 H}{\partial \vec{m}_A(\vec{r}_A) \partial \vec{m}_B(\vec{r}'_B)} \cdot \delta \vec{m}_B(\vec{r}'_B) d^3 \vec{r}_A d^3 \vec{r}'_B \quad (67)$$

is greater than or equal to zero for arbitrary $\{\delta \vec{m}_A(\vec{r}_A)\}$. Since the only non-zero second order derivatives of H with respect to the independent variables $\{\vec{m}_A(\vec{r}_A)\}$ are

$$\frac{\partial^2 H}{\partial \vec{m}_A(\vec{r}_A) \partial \vec{m}_A(\vec{r}'_A)} = \frac{1}{2} \oint_{\omega} \frac{\partial}{\partial \vec{m}_A(\vec{r}'_A)} \ln \left(\frac{\rho_A(\vec{r}_A, \hat{h})}{w_A^{\text{spin}}(\vec{r}_A, \hat{h})} \right) \hat{h} d^2 \omega + \nu_A(\vec{r}_A) \frac{\partial \hat{m}_A(\vec{r}_A)}{\partial \vec{m}_A(\vec{r}'_A)} \quad (68)$$

it follows from eq 67 that H is convex if

$$\oint \oint \vec{\varphi}(\vec{r}_A) \cdot \frac{\partial^2 H}{\partial \vec{m}_A(\vec{r}_A) \partial \vec{m}_A(\vec{r}'_A)}(\vec{r}'_A) \cdot \vec{\varphi}(\vec{r}'_A) d^3 \vec{r}_A d^3 \vec{r}'_A \geq 0 \quad (69)$$

for every real-valued differentiable function $\vec{\varphi}(\vec{r}_A)$. Expanding,

$$\frac{\partial \hat{m}_A(\vec{r}_A)}{\partial \vec{m}_A(\vec{r}'_A)} = \left(\frac{\hat{x}\hat{x} + \hat{y}\hat{y} + \hat{z}\hat{z} - \hat{m}_A(\vec{r}_A)\hat{m}_A(\vec{r}'_A)}{m_A(\vec{r}_A)} \right) \delta^{\text{dirac}}(\vec{r}'_A - \vec{r}_A) \quad (70)$$

gives

$$\begin{aligned} & \oint \oint \vec{\varphi}(\vec{r}_A) \cdot \nu_A(\vec{r}_A) \frac{\partial \hat{m}_A(\vec{r}_A)}{\partial \vec{m}_A(\vec{r}'_A)} \cdot \vec{\varphi}(\vec{r}'_A) d^3 \vec{r}_A d^3 \vec{r}'_A \\ &= \oint \nu_A(\vec{r}_A) \frac{|\vec{\varphi}(\vec{r}_A)|^2 - |\vec{\varphi}(\vec{r}_A) \cdot \hat{m}_A(\vec{r}_A)|^2}{m_A(\vec{r}_A)} d^3 \vec{r}_A \geq 0 \end{aligned} \quad (71)$$

The inequality in eq 71 follows from the fact that $\nu_A(\vec{r}_A) \geq 0$ and the absolute value of the scalar projection of a vector onto any unit direction is less than or equal to the magnitude of that vector. Expanding

$$\begin{aligned} & \frac{\partial}{\partial \vec{m}_A(\vec{r}'_A)} \ln \left(\frac{\rho_A(\vec{r}_A, \hat{h})}{w_A^{\text{spin}}(\vec{r}_A, \hat{h})} \right) \\ &= \hat{h} \left(\frac{\delta^{\text{dirac}}(\vec{r}'_A - \vec{r}_A)}{2\rho_A(\vec{r}_A, \hat{h})} - \frac{(1 - \chi_{\text{spin}})\delta^{\text{dirac}}(\vec{r}'_A - \vec{r}_A)}{8\pi(r_A)^2 \rho_A^{\text{avg}}(r_A, \hat{h})} \right) \end{aligned} \quad (72)$$

gives

$$\begin{aligned} & \oint \oint \vec{\varphi}(\vec{r}_A) \cdot \hat{h} \frac{\partial}{\partial \vec{m}_A(\vec{r}'_A)} \ln \left(\frac{\rho_A(\vec{r}_A, \hat{h})}{w_A^{\text{spin}}(\vec{r}_A, \hat{h})} \right) \hat{h} \vec{\varphi}(\vec{r}'_A) d^3 \vec{r}_A d^3 \vec{r}'_A \\ &= \oint \frac{|\vec{\varphi}(\vec{r}_A) \cdot \hat{h}|^2}{2\rho_A(\vec{r}_A, \hat{h})} d^3 \vec{r}_A - (1 - \chi_{\text{spin}}) \oint \frac{|\vec{\varphi}(\vec{r}_A) \cdot \hat{h}| |\vec{\varphi}^{\text{avg}}(r_A) \cdot \hat{h}|}{2\rho_A^{\text{avg}}(r_A, \hat{h})} d^3 \vec{r}_A \\ &= \oint \left(\frac{|\vec{\varphi}(\vec{r}_A) \cdot \hat{h}|^2}{2\rho_A(\vec{r}_A, \hat{h})} - (1 - \chi_{\text{spin}}) \frac{|\vec{\varphi}^{\text{avg}}(r_A) \cdot \hat{h}|^2}{2\rho_A^{\text{avg}}(r_A, \hat{h})} \right) d^3 \vec{r}_A \geq 0 \end{aligned} \quad (73)$$

Bultinck et al. showed that

$$\oint \left(\frac{f(\vec{r}_A)^2}{g_A(\vec{r}_A)} - \frac{f^{\text{avg}}(r_A)^2}{g_A^{\text{avg}}(r_A)} \right) d^3 \vec{r}_A \geq 0 \quad (74)$$

for any differentiable real-valued functions $g_A(\vec{r}_A) \geq 0$ and $f(\vec{r}_A)$.¹²⁵ Thus, by setting $f(\vec{r}_A) = \vec{\varphi}(\vec{r}_A) \cdot \hat{h}$ and $g_A(\vec{r}_A) = 2\rho_A(\vec{r}_A, \hat{h})$, it follows that the integral in eq 73 is positive definite. Combining eqs 67, 68, 71, and 73 gives $\delta^2 H \geq 0$, which proves that H has a unique minimum.

2. Algorithm for Computing the Inverse of $\xi(\vec{r})$. The iterative scheme defined in section 3.1 requires calculating the

inverse of $\gamma = \xi(\tau)$, denoted $\xi^{\text{inv}}(\gamma)$. This quantity was computed by Newton's method starting from an initial estimate of

$$\xi^{\text{inv}}(\gamma)|_0 = \frac{\gamma + a_1\gamma^2}{a_2 + a_3\gamma + a_4\gamma^2}, 0 \leq \gamma \leq 2\pi \quad (75)$$

The constants $a_1 = -0.036835$, $a_2 = 4.178319$, $a_3 = -0.136129$, and $a_4 = 0.038148$ were determined by fitting the points $\gamma = 0$, $\pi/2$, π , $3\pi/2$, and 2π . Newton's method gives the refined value at iteration ($j+1$) in terms of the error and derivative at iteration j ; specifically,

$$\xi^{\text{inv}}(\gamma)|_{j+1} = \xi^{\text{inv}}(\gamma)|_j + \frac{\gamma - \xi(\xi^{\text{inv}}(\gamma)|_j)}{\xi'(\xi^{\text{inv}}(\gamma)|_j)} \quad (76)$$

where the derivative is

$$\xi'(\tau) = \frac{2\pi}{\tau^3} \left(\ln \left(\frac{1+\tau}{1-\tau} \right) - 2\tau \right), \xi'(0) = 4\pi/3 \quad (77)$$

$\xi(\tau)$ was expanded near $\tau = 0$ and 1 to avoid singularities in eqs 35 and 77: (a) $\xi(\tau) \approx 4\pi\tau/3$ and $\xi'(\tau) \approx 4\pi/3$ for $\tau \leq 0.001$ and (b) $\xi(\tau) \approx 2\pi + 49.956(\tau - 1) + 8370.6(\tau - 1)^2$ and $\xi'(\tau) \approx 50.0$ for $\tau \geq 0.999$. Two Newton iterations were sufficient to achieve at least four significant digits of accuracy over the entire range of τ values.

3. Derivation of the Value $c = \pi$ in eq 59. Since eq 64 imposes constraint 11 for any positive value of c , the value of c has no effect on the converged solution. It does, however, affect the number of iterations required to reach convergence. To minimize the number of iterations required to reach convergence, the value $c = \pi$ derived here should be used. Using the subscripts j and $(j-1)$ to indicate assignments for the current iteration (i.e., j) based on quantities for the last iteration (i.e., $j-1$), eq 59 is

$$\begin{aligned} \bar{L}_A(\vec{r}_A)|_j &= \bar{Y}(\vec{r})|_{j-1} - \bar{Q}(\vec{r})|_{j-1} \\ &+ c \frac{(\bar{m}(\vec{r}) - \bar{m}^{\text{trial}}(\vec{r}))}{\rho(\vec{r})} \Big|_{j-1} + \bar{\eta}_A(\vec{r}_A)|_{j-1} \end{aligned} \quad (78)$$

Multiplying both sides by $\rho_A(\vec{r}_A)/\rho(\vec{r})$ and summing over A gives

$$\bar{Y}(\vec{r})|_j - \bar{Y}(\vec{r})|_{j-1} = c \frac{(\bar{m}(\vec{r}) - \bar{m}^{\text{trial}}(\vec{r}))|_{j-1}}{\rho(\vec{r})} \quad (79)$$

To satisfy constraint 11 in iteration j , we should have

$$\bar{m}^{\text{trial}}(\vec{r})|_j \approx \bar{m}(\vec{r}) \quad (80)$$

Inserting eq 80 into the right-hand side of eq 79 gives

$$\bar{Y}(\vec{r})|_j - \bar{Y}(\vec{r})|_{j-1} \approx \frac{c}{\rho(\vec{r})} (\bar{m}^{\text{trial}}(\vec{r})|_j - \bar{m}^{\text{trial}}(\vec{r})|_{j-1}) \quad (81)$$

Defining

$$\Delta m_A(\vec{r}_A) = m_A(\vec{r}_A)|_j - m_A(\vec{r}_A)|_{j-1} \quad (82)$$

Equation 81 can be rewritten as

$$\frac{\partial \bar{Y}(\vec{r})}{\partial m_A(\vec{r}_A)} \Delta m_A(\vec{r}_A) \approx \frac{c}{\rho(\vec{r})} \Delta m_A(\vec{r}_A) \quad (83)$$

Using eqs 61 and 37, the derivative can be expanded as

$$\frac{\partial \bar{Y}(\vec{r})}{\partial m_A(\vec{r}_A)} = \frac{\rho_A(\vec{r}_A)}{\rho(\vec{r})} \frac{\partial L_A(\vec{r}_A)}{\partial m_A(\vec{r}_A)} = \frac{1}{2\rho(\vec{r})} \frac{\partial \xi(\tau)}{\partial \tau} \quad (84)$$

Comparing eqs 83 and 84, one-half the average value of the derivative of $\xi(\tau)$ over the range of possible τ values (i.e. $\tau = 0$ to 1) is a good choice for c :

$$c \approx \frac{1}{2} \frac{\partial \xi(\tau)}{\partial \tau} \approx \frac{\xi(1) - \xi(0)}{2} = \frac{2\pi - 0}{2} = \pi \quad (85)$$

In practice, we found that using $c = \pi$ converges $\bar{m}^{\text{trial}}(\vec{r})$ to $\bar{m}(\vec{r})$ in just a couple of iterations.

■ ASSOCIATED CONTENT

S Supporting Information. Coordinates of optimized geometries, more extensive results tables, example input files, set of grid points used to compute the MAE and RMAE of $\bar{B}_{\text{spin}}(\vec{r})$, and computational algorithm and uniqueness proof for the LSF functional that computes ASMs by minimizing $|\bar{m}_A(\vec{r}_A) - \bar{m}_A^{\text{avg}}(r_A)|^2 / \rho_A^{\text{avg}}(r_A)$ subject to constraints 11 and 26. This material is available free of charge via the Internet at <http://pubs.acs.org>.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: thomasamanz@gmail.com, david.sholl@chbe.gatech.edu.

■ ACKNOWLEDGMENT

This material is based on work supported as part of the Center for Atomic Level Catalyst Design, an Energy Frontier Research Center funded by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences under Award Number DE-SC0001058 (Center for Atomic Level Catalyst Design). Supercomputing resources were provided by the Georgia Institute of Technology and Teragrid grant TG-CTS100027.

■ REFERENCES

- (1) Back, C. H.; Wursch, C.; Vaterlaus, A.; Ramsperger, U.; Maier, U.; Pescia, D. *Nature* **1995**, *378*, 597–600.
- (2) Rosenbaum, T. F.; Silevitch, D. M.; Bitko, D.; Brooke, J.; Ghosh, S.; Aeppli, G. *Nature* **2007**, *448*, 567–570.
- (3) Zhang, W. M.; Saslow, W. M.; Gabay, M.; Benakli, M. *Phys. Rev. B* **1993**, *48*, 10204–10216.
- (4) Bader, R. F. W. *J. Phys. Chem. A* **2007**, *111*, 7966–7972.
- (5) Bader, R. F. W.; Macdougall, P. J.; Lau, C. D. H. *J. Am. Chem. Soc.* **1984**, *106*, 1594–1605.
- (6) Hirshfeld, F. L. *Theor. Chim. Acta* **1977**, *44*, 129–138.
- (7) Reed, A. E.; Weinstock, R. B.; Weinhold, F. *J. Chem. Phys.* **1985**, *83*, 735–746.
- (8) Philips, J. J.; Hudspeth, M. A.; Browne, P. M.; Peralta, J. E. *Chem. Phys. Lett.* **2010**, *495*, 146–150.
- (9) Macdougall, P. J.; Bader, R. F. W. *Can. J. Chem.* **1986**, *64*, 1496–1508.
- (10) Mulliken, R. S. *J. Chem. Phys.* **1955**, *23*, 1833–1840.

- (11) Kresse, G.; Marsman, M.; Furthmüller, J. *VASP the Guide*; Universität Wien: Vienna, Austria, 2011. <http://cms.mpi.univie.ac.at/vasp/vasp/vasp.html> (accessed July 26, 2011).
- (12) Bultinck, P.; Van Alsenoy, C.; Ayers, P. W.; Carbo-Dorca, R. *J. Chem. Phys.* **2007**, *126*, 144111.
- (13) Davidson, E. R.; Chakravorty, S. *Theor. Chim. Acta* **1992**, *83*, 319–330.
- (14) Manz, T. A.; Sholl, D. S. *J. Chem. Theory Comput.* **2010**, *6*, 2455–2468.
- (15) Montgomery, J. A.; Frisch, M. J.; Ochterski, J. W.; Petersson, G. A. *J. Chem. Phys.* **2000**, *112*, 6532–6542.
- (16) Watanabe, T.; Manz, T. A.; Sholl, D. S. *J. Phys. Chem. C* **2011**, *115*, 4824–4836.
- (17) Kim, H. W.; Jung, J.; Han, M. N.; Lim, S.; Tamada, K.; Hara, M.; Kawa, M.; Kim, Y.; Kuk, Y. *J. Am. Chem. Soc.* **2011**, *133*, 9236–9238.
- (18) Dirac, P. A. M. *Proc. R. Soc., London A* **1928**, *118*, 351–361.
- (19) Dirac, P. A. M. *Proc. R. Soc., London A* **1928**, *117*, 610–624.
- (20) Foldy, L. L.; Wouthuysen, S. A. *Phys. Rev.* **1950**, *78*, 345–345.
- (21) Szabo, A.; Ostlund, N. S. *Modern Quantum Chemistry*; Dover: Mineola, NY, 1996; pp 97–107.
- (22) Peralta, J. E.; Scuseria, G. E.; Frisch, M. J. *Phys. Rev. B* **2007**, *75*, 125119.
- (23) van Lenthe, E.; Snijders, J. G.; Baerends, E. J. *J. Chem. Phys.* **1996**, *105*, 6505–6516.
- (24) Closs, G. L.; Forbes, M. D. E.; Piotrowiak, P. *J. Am. Chem. Soc.* **1992**, *114*, 3285–3294.
- (25) Gidopoulos, N. I. *Phys. Rev. B* **2007**, *75*, 134408.
- (26) Ivanovic, I. D. *J. Phys. A* **1993**, *26*, L579–L582.
- (27) Wittbrodt, J. M.; Schlegel, H. B. *J. Chem. Phys.* **1996**, *105*, 6574–6577.
- (28) Cohen, A. J.; Tozer, D. J.; Handy, N. C. *J. Chem. Phys.* **2007**, *126*, 214104.
- (29) Nakatsuji, S.; Machida, Y.; Onoda, S.; Tayama, T.; Sakakibara, T. *Nature* **2010**, *463*, 210–213.
- (30) Evans, P. G.; Isaacs, E. D.; Aeppli, G.; Cai, Z.; Lai, B. *Science* **2002**, *295*, 1042–1045.
- (31) Gingras, M. J. P.; Bramwell, S. T. *Science* **2001**, *294*, 1495–1501.
- (32) Lee, S. H.; Broholm, C.; Ratcliff, W.; Gasparovic, G.; Huang, Q.; Kim, T. H.; Cheong, S. W. *Nature* **2002**, *418*, 856–858.
- (33) Shpyrko, O. G.; Isaacs, E. D.; Logan, J. M.; Feng, Y. J.; Aeppli, G.; Jaramillo, R.; Kim, H. C.; Rosenbaum, T. F.; Zschack, P.; Sprung, M.; Narayanan, S.; Sandy, A. R. *Nature* **2007**, *447*, 68–71.
- (34) Perdew, J. P.; Ruzsinszky, A.; Constantin, L. A.; Sun, J. W.; Csonka, G. I. *J. Chem. Theory Comput.* **2009**, *5*, 902–908.
- (35) Barth, U. V.; Hedin, L. *J. Phys. C* **1972**, *5*, 1629–1642.
- (36) Pople, J. A.; Gill, P. M. W.; Handy, N. C. *Int. J. Quantum Chem.* **1995**, *56*, 303–305.
- (37) Grafenstein, J.; Kraka, E.; Cremer, D. *Chem. Phys. Lett.* **1998**, *288*, 593–602.
- (38) Grafenstein, J.; Kraka, E.; Filatov, M.; Cremer, D. *Int. J. Mol. Sci.* **2002**, *3*, 360–394.
- (39) Perdew, J. P.; Savin, A.; Burke, K. *Phys. Rev. A* **1995**, *51*, 4531–4541.
- (40) Caballol, R.; Castell, O.; Illas, F.; Moreira, P. R.; Malrieu, J. P. *J. Phys. Chem. A* **1997**, *101*, 7860–7866.
- (41) Grimme, S.; Waletzke, M. *J. Chem. Phys.* **1999**, *111*, 5645–5655.
- (42) Ullrich, C. A.; Kohn, W. *Phys. Rev. Lett.* **2001**, *87*, 093001.
- (43) van Leeuwen, R. *Adv. Quantum Chem.* **2003**, *43*, 25–94.
- (44) Dreizler, R. M.; Gross, E. K. U. *Density Functional Theory*; Springer: Berlin, 1990; pp 54–61.
- (45) Nakatsuji, H. *Chem. Phys. Lett.* **1979**, *67*, 329–333.
- (46) Nakatsuji, H.; Hirao, K. *J. Chem. Phys.* **1978**, *68*, 2053–2065.
- (47) Ehara, M.; Kuramoto, K.; Nakatsuji, H. *Chem. Phys.* **2009**, *356*, 195–198.
- (48) Geertsen, J.; Rittby, M.; Bartlett, R. J. *Chem. Phys. Lett.* **1989**, *164*, 57–62.
- (49) Evangelista, F. A. *J. Chem. Phys.* **2011**, *134*, 224102.
- (50) Ruiz, E.; Cirera, J.; Alvarez, S. *Coord. Chem. Rev.* **2005**, *249*, 2649–2660.
- (51) Lillestolen, T. C.; Wheatley, R. J. *Chem. Commun.* **2008**, 5909–5911.
- (52) Good, R. H.; Nelson, T. J. *Classical Theory of Electric and Magnetic Fields*; Academic Press: New York, 1971, pp 219–220, 465–466.
- (53) Panofsky, W. K. H.; Phillips, M. *Classical Electricity and Magnetism*; Addison-Wesley: Reading, MA, 1962; pp 130–133.
- (54) Campana, C.; Mussard, B.; Woo, T. K. *J. Chem. Theory Comput.* **2009**, *5*, 2866–2878.
- (55) Capelle, K.; Vignale, G.; Gyroff, B. L. *Phys. Rev. Lett.* **2001**, *87*, 206403.
- (56) Sharma, S.; Dewhurst, J. K.; Ambrosch-Draxl, C.; Kurth, S.; Helbig, N.; Pittalis, S.; Shallcross, S.; Nordstrom, L.; Gross, E. K. U. *Phys. Rev. Lett.* **2007**, *98*, 196405.
- (57) Hobbs, D.; Kresse, G.; Hafner, J. *Phys. Rev. B* **2000**, *62*, 11556–11570.
- (58) Oda, T.; Pasquarello, A.; Car, R. *Phys. Rev. Lett.* **1998**, *80*, 3622–3625.
- (59) Knopfle, K.; Sandratskii, L. M.; Kubler, J. *Phys. Rev. B* **2000**, *62*, 5564–5569.
- (60) Tsunoda, Y.; Nishioka, Y.; Nicklow, R. M. *J. Magn. Magn. Mater.* **1993**, *128*, 133–137.
- (61) Hafner, J. *J. Comput. Chem.* **2008**, *29*, 2044–2078.
- (62) Kresse, G.; Furthmüller, J. *Phys. Rev. B* **1996**, *54*, 11169–11186.
- (63) Kresse, G.; Joubert, D. *Phys. Rev. B* **1999**, *59*, 1758–1775.
- (64) 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, O.; Foresman, J. B.; Ortiz, J. V.; Ciolowski, J.; Fox, D. *J. Gaussian 09, Revision A.02*; Gaussian, Inc.: Wallingford, CT, 2009.
- (65) Watanabe, T.; Sholl, D. S. *J. Chem. Phys.* **2010**, *133*, 094509.
- (66) Zuxiang, Y. *Rock Miner. Anal.* **1984**, *3*, 231–238.
- (67) Louie, S. G.; Froyen, S.; Cohen, M. L. *Phys. Rev. B* **1982**, *26*, 1738–1742.
- (68) Krauss, M.; Stevens, W. J. *Annu. Rev. Phys. Chem.* **1984**, *35*, 357–385.
- (69) Henkelman, G.; Arnaldsson, A.; Jonsson, H. *Comput. Mater. Sci.* **2006**, *36*, 354–360.
- (70) Sanville, E.; Kenny, S. D.; Smith, R.; Henkelman, G. *J. Comput. Chem.* **2007**, *28*, 899–908.
- (71) Tang, W.; Sanville, E.; Henkelman, G. *J. Phys. Cond. Matter* **2009**, *21*, 084204.
- (72) Cao, W. L.; Gatti, C.; Macdougall, P. J.; Bader, R. F. W. *Chem. Phys. Lett.* **1987**, *141*, 380–385.
- (73) Glendening, E. D.; Reed, A. E.; Carpenter, J. E.; Weinhold, F. *NBO 3.0 Program Manual*; University of Wisconsin: Madison, WI, 1990; pp A.1–B.56.
- (74) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865–3868.
- (75) Singh, U. C.; Kollman, P. A. *J. Comput. Chem.* **1984**, *5*, 129–145.
- (76) Aebbersold, M. A.; Gillon, B.; Plantevin, O.; Pardi, L.; Kahn, O.; Bergerat, P.; von Seggern, I.; Tuczek, F.; Ohrstrom, L.; Grand, A.; Lelievre-Berna, E. *J. Am. Chem. Soc.* **1998**, *120*, 5238–5245.
- (77) Figgis, B. N.; Forsyth, J. B.; Reynolds, P. A. *Inorg. Chem.* **1987**, *26*, 101–105.
- (78) Hines, W. A.; Menotti, A. H.; Budnick, J. I.; Burch, T. J.; Litrenta, T.; Niculescu, V.; Raj, K. *Phys. Rev. B* **1976**, *13*, 4060–4068.

- (79) Wenzel, M. J.; Steinle-Neumann, G. *Phys. Rev. B* **2007**, *75*, 214430.
- (80) Verwey, E. J. W. *Nature* **1939**, *144*, 327–328.
- (81) Shchennikov, V. V.; Ovsyannikov, S. V. *J. Phys.: Condens. Matter* **2009**, *21*, 271001.
- (82) Pinto, H. P.; Elliott, S. D. *J. Phys.: Condens. Matter* **2006**, *18*, 10427–10436.
- (83) Mantovan, R.; Lamperti, A.; Georgieva, M.; Tallarida, G.; Fanciulli, M. *J. Phys. D* **2010**, *43*, 065002.
- (84) Zhang, M.; Yan, H.; Yan, H. *J. Magn. Magn. Mater.* **2009**, *321*, 2340–2344.
- (85) Zhang, D. H.; Liu, Z. Q.; Han, S.; Li, C.; Lei, B.; Stewart, M. P.; Tour, J. M.; Zhou, C. W. *Nano Lett.* **2004**, *4*, 2151–2155.
- (86) Yoon, K. S.; Koo, J. H.; Do, Y. H.; Kim, K. W.; Kim, C. O.; Hong, J. P. *J. Magn. Magn. Mater.* **2005**, *285*, 125–129.
- (87) Parkinson, G. S.; Mulakaluri, N.; Losovyj, Y.; Jacobson, P.; Pentcheva, R.; Diebold, U. *Phys. Rev. B* **2010**, *82*, 125413.
- (88) Kurahashi, M.; Sun, X.; Yamauchi, Y. *Phys. Rev. B* **2010**, *81*, 193402.
- (89) Lodziana, Z. *Phys. Rev. Lett.* **2007**, *99*, 206402.
- (90) Pentcheva, R.; Wendler, F.; Meyerheim, H. L.; Moritz, W.; Jedrecy, N.; Scheffler, M. *Phys. Rev. Lett.* **2005**, *94*, 126101.
- (91) Dudarev, S. L.; Botton, G. A.; Savrasov, S. Y.; Humphreys, C. J.; Sutton, A. P. *Phys. Rev. B* **1998**, *57*, 1505–1509.
- (92) Rakhecha, V. C.; Murthy, N. S. S. *J. Phys. C* **1978**, *11*, 4389–4404.
- (93) Goering, E.; Gold, S.; Lafkioti, M.; Schutz, G. *Europhys. Lett.* **2006**, *73*, 97–103.
- (94) Heaton-Burgess, T.; Yang, W. T. *J. Chem. Phys.* **2010**, *132*, 234113.
- (95) Mori-Sanchez, P.; Cohen, A. J.; Yang, W. T. *J. Chem. Phys.* **2006**, *125*, 201102.
- (96) Perdew, J. P.; Parr, R. G.; Levy, M.; Balduz, J. L. *Phys. Rev. Lett.* **1982**, *49*, 1691–1694.
- (97) Mori-Sanchez, P.; Cohen, A. J.; Yang, W. T. *Phys. Rev. Lett.* **2008**, *100*, 146401.
- (98) Chevrier, V. L.; Ong, S. P.; Armiento, R.; Chan, M. K. Y.; Ceder, G. *Phys. Rev. B* **2010**, *82*, 075122.
- (99) van der Marel, D.; Sawatzky, G. A. *Phys. Rev. B* **1988**, *37*, 10674–10684.
- (100) Cohen, A. J.; Mori-Sanchez, P.; Yang, W. T. *J. Chem. Phys.* **2007**, *126*, 191109.
- (101) Krukau, A. V.; Vydrov, O. A.; Izmaylov, A. F.; Scuseria, G. E. *J. Chem. Phys.* **2006**, *125*, 224106.
- (102) Paier, J.; Marsman, M.; Hummer, K.; Kresse, G.; Gerber, I. C.; Angyan, J. G. *J. Chem. Phys.* **2006**, *125*, 249901.
- (103) Paier, J.; Marsman, M.; Hummer, K.; Kresse, G.; Gerber, I. C.; Angyan, J. G. *J. Chem. Phys.* **2006**, *124*, 154709.
- (104) Oshio, H.; Hoshino, N.; Ito, T.; Nakano, M. *J. Am. Chem. Soc.* **2004**, *126*, 8805–8812.
- (105) Oshio, H.; Hoshino, N.; Ito, T. *J. Am. Chem. Soc.* **2000**, *122*, 12602–12603.
- (106) Vosko, S. H.; Wilk, L.; Nusair, M. *Can. J. Phys.* **1980**, *58*, 1200–1211.
- (107) Perdew, J. P.; Chevary, J. A.; Vosko, S. H.; Jackson, K. A.; Pederson, M. R.; Singh, D. J.; Fiolhais, C. *Phys. Rev. B* **1992**, *46*, 6671–6687.
- (108) Bogani, L.; Wernsdorfer, W. *Nat. Mater.* **2008**, *7*, 179–186.
- (109) Oshio, H.; Nakano, M. *Chem.—Eur. J.* **2005**, *11*, 5178–5185.
- (110) Ribas-Arino, J.; Baruah, T.; Pederson, M. R. *J. Am. Chem. Soc.* **2006**, *128*, 9497–9505.
- (111) De Vos, D. E.; Alaerts, L.; Seguin, E.; Poelman, H.; Thibault-Starzyk, F.; Jacobs, P. A. *Chem.—Eur. J.* **2006**, *12*, 7353–7363.
- (112) Walton, K. S.; Karra, J. R. *Langmuir* **2008**, *24*, 8620–8626.
- (113) Prestipino, C.; Regli, L.; Vitillo, J. G.; Bonino, F.; Damin, A.; Lamberti, C.; Zecchina, A.; Solari, P. L.; Kongshaug, K. O.; Bordiga, S. *Chem. Mater.* **2006**, *18*, 1337–1346.
- (114) Schlichte, K.; Kratzke, T.; Kaskel, S. *Microporous Mesoporous Mater.* **2004**, *73*, 81–88.
- (115) Thomas, K. M.; Xiao, B.; Wheatley, P. S.; Zhao, X. B.; Fletcher, A. J.; Fox, S.; Rossi, A. G.; Megson, I. L.; Bordiga, S.; Regli, L.; Morris, R. E. *J. Am. Chem. Soc.* **2007**, *129*, 1203–1209.
- (116) Liu, Y.; Xiang, S. C.; Zhou, W.; Gallegos, J. M.; Chen, B. L. *J. Am. Chem. Soc.* **2009**, *131*, 12415–12419.
- (117) Poppl, A.; Kunz, S.; Himsl, D.; Hartmann, M. *J. Phys. Chem. C* **2008**, *112*, 2678–2684.
- (118) Zhang, X. X.; Chui, S. S. Y.; Williams, I. D. *J. Appl. Phys.* **2000**, *87*, 6007–6009.
- (119) Bohlmann, W.; Poppl, A.; Sabo, M.; Kaskel, S. *J. Phys. Chem. B* **2006**, *110*, 20177–20181.
- (120) Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H. *SAC/SAC-CI Program Combined with Gaussian for Calculating Ground, Excited, Ionized, and Electron-Attached States and Singlet, Doublet, Triplet, Quartet, Quintet, Sextet, and Septet Spin States and Their Analytic Energy Gradients*; Kyoto University: Kyoto, Japan, 2005; pp 1–450.
- (121) Tyuterev, V. G.; Tashkun, S.; Jensen, P.; Barbe, A.; Cours, T. *J. Mol. Spectrosc.* **1999**, *198*, 57–76.
- (122) Bacis, R.; Bouvier, A. J.; Flaud, J. M. *Spectrochim. Acta, Part A* **1998**, *54*, 17–34.
- (123) Bouvier, A. J.; Inard, D.; Veyret, V.; Bussery, B.; Bacis, R.; Churassy, S.; Brion, J.; Malicet, J.; Judge, R. H. *J. Mol. Spectrosc.* **1998**, *190*, 189–197.
- (124) Willitsch, S.; Innocenti, F.; Dyke, J. M.; Merkt, F. *J. Chem. Phys.* **2005**, *122*, 024311.
- (125) Bultinck, P.; Cooper, D. L.; Van Neck, D. *Phys. Chem. Chem. Phys.* **2009**, *11*, 3424–3429.

Structural Anisotropy in Polar Fluids Subjected to Periodic Boundary Conditions

Joakim Stenhammar,^{*,†} Gunnar Karlström,[‡] and Per Linse[†][†]Division of Physical Chemistry, Lund University, Sweden[‡]Division of Theoretical Chemistry, Lund University, Sweden

ABSTRACT: A heuristic model based on dielectric continuum theory for the long-range solvation free energy of a dipolar system possessing periodic boundary conditions (PBCs) is presented. The predictions of the model are compared to simulation results for Stockmayer fluids simulated using three different cell geometries. The boundary effects induced by the PBCs are shown to lead to anisotropies in the apparent dielectric constant and the long-range solvation free energy of as much as 50%. However, the sum of all of the anisotropic energy contributions yields a value that is very close to the isotropic one derived from dielectric continuum theory, leading to a total system energy close to the dielectric value. It is finally shown that the leading-order contribution to the energetic and structural anisotropy is significantly smaller in the noncubic simulation cell geometries compared to when using a cubic simulation cell.

1. INTRODUCTION

Molecular dynamics (MD) and Monte Carlo (MC) simulations have grown to become a central tool in physics, chemistry, and biology over the past three decades.^{1,2} However, in spite of the huge advancement of both algorithms and hardware, there are still some unresolved methodological issues. Arguably, the most persistent of these is the question of how to handle long-range electrostatic (Coulomb and dipole–dipole) interactions in a simulation.^{3–5} The basic problem is that the integral

$$\int_{r_{\text{cut}}}^{\infty} v(r) 4\pi r^2 dr \quad (1)$$

diverges for all finite values of the cutoff radius r_{cut} as long as the intermolecular potential $v(r)$ does not decay faster than r^{-3} . Thus, applying a simple (spherical or cubic) cutoff to the electrostatic potentials may often lead to serious artifacts in the structure and thermodynamics of the system under study.

Although several solutions to the infinite-range interaction problem have been proposed, the most common way to circumvent this problem is the use of lattice-based summation techniques, or *periodic boundary conditions* (PBCs). These methods compose a plethora of different algorithms that all rest on the same basic assumption, namely that the (finitely sized) simulation cell is duplicated in all directions to create an infinite lattice. The original implementation of this idea was developed by Ewald⁶ and is built upon a separation of the interaction into short-range and long-range parts, where the former is summed up in real space and the latter in reciprocal space. The original Ewald method has since been developed in many ways, and today different mesh-based methods^{7–10} are numerically faster alternatives to the classical Ewald summation.

When simulating a fluid phase, the assumption of periodicity is clearly not a correct description of the real system. This criticism has been put forward several times in the literature but was originally noted by Valleau and Whittington,¹¹ who gave a qualitative argument about the inability of lattice summation

methods to correctly reproduce long-range fluctuations in fluid systems. Furthermore, several studies have addressed the issue of periodicity effects on the properties of Lennard-Jones fluids,^{12,13} ionic solutions,^{14–17} and biomolecules.^{18–21} In the context of dipolar systems, Boresch and Steinhauser²² conducted a careful study of dipole fluctuations and correlations in SPC water simulated using the Ewald summation technique. In particular, they addressed the importance of the so-called surface term,²³ which describes the solvation from the dielectric surroundings of the infinite lattice on structural properties such as the dielectric permittivity, dipole time correlation functions, and the Kirkwood g factor. However, the total dipole moment of the simulation box is a special property, in the sense that its total interaction with all its periodic images is identically zero, as long as the contributions are summed in spherical shells.^{24–27} Therefore, the periodicity effects on the fluctuating dipole moment of the whole simulation box (and related properties) are expected to be small. In a recent contribution,²⁸ we showed, however, that the fluctuations of higher order electric multipole moments of the whole simulation box are greatly influenced by the interaction between each instantaneous multipole and all of its periodic images. This effect is manifested through a difference of as much as 50% between the dielectric permittivities calculated from different multipole components, depending on whether the multipole component has an attractive or a repulsive (or, in some cases, zero) interaction with its neighbors. A schematic picture of the coupling of different multipole components in a system under PBCs is given in Figure 1.

In addition to the cubic simulation cell used in the majority of computer simulations, some alternative simulation cell geometries have been suggested and implemented,^{29–33} most notably the rhombic dodecahedron (RD) and the truncated octahedron (TO). These two bodies have the appealing property of more

Received: August 22, 2011

Published: October 18, 2011

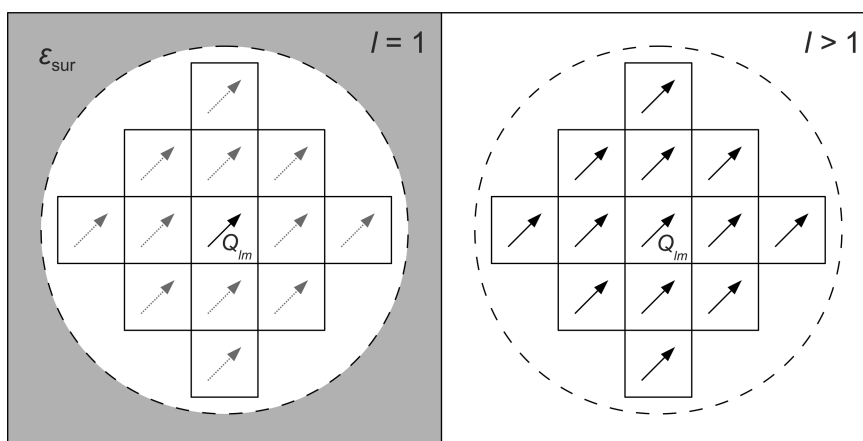


Figure 1. Schematic picture of the coupling of the total dipole (left) and higher multipoles (right) of a simulation cell subjected to PBCs. The dipole does not “see” its neighbors since its self-interaction energy is zero but is solvated by the dielectric response from the surrounding medium through the surface term. In contrast, higher multipoles (Q_{lm} , $l > 1$) couple to its neighbors through their nonzero self-interaction but are not affected by the surface term. In addition, the dipole as well as the higher multipoles interact with the set of “unconstrained” multipoles $Q_{l'm'}$, $l' \neq l$ and/or $m' \neq m$, (not depicted) in the surrounding cells, giving an (approximately) isotropic contribution to the solvation.

closely resembling the geometry of solvated spherical solutes, in the sense that they have larger inscribed spheres than a cubic simulation cell of the same volume. Even though these alternative geometries are implemented in major simulation packages, there are only a few studies^{29,34,35} probing the effect of changing the cell geometry on the thermodynamic properties of the system under study. Because these cells pack in lattice structures different from that of the cube, it seems reasonable to expect their periodicity effects to differ qualitatively from those of a cubic cell.

In the present contribution, we will extend our previous analysis²⁸ of the periodicity effects on a dipolar model system from the qualitative to the quantitative level, as well as from cubic to noncubic simulation cells. We will develop a heuristic model describing the solvation of an electrostatic fluctuation in a spherical subvolume of a dielectric medium exhibited to PBCs. This model will be compared to values of the dielectric constant calculated from simulations of a simple dipolar model system.

2. THEORY

2.1. General ansatz. In the following, the electrostatic fluctuations in a spherical subvolume of a dipolar model system treated using the Ewald summation technique will be described by dividing the long-range solvation energy of this subvolume into two contributions:

- An approximately isotropic part, coming from the interaction between the instantaneous multipole Q_{lm} of a spherical volume in the central simulation cell and its noncorrelated neighbors, i.e., $Q_{l'm'}$ with $l' \neq l$ and/or $m' \neq m$, in the other cells. This interaction is, at least partly, Boltzmann-weighted in a simulation, and we will thus attempt to describe it using formulas valid for an isotropic dielectric medium.
- A strongly anisotropic part, coming from the “self-interaction” between Q_{lm} in the central cell and its fully correlated replicas ($Q_{l'm'}$ with $l' = l$ and $m' = m$) in the rest of the lattice. This part of the interaction is not Boltzmann-weighted, because of the perfect periodicity imposed by the PBCs. We

will thus describe this interaction using the reduced lattice-interaction tensors introduced previously.²⁸

On the basis of this description, we will present a heuristic derivation of the long-range solvation free energy of the spherical subvolume. This will be compared to the behavior expected from a spherical subvolume inside an infinite isotropic dielectric medium and the analysis will thus enable us to directly probe the magnitude of the periodicity effect introduced by the PBCs.

2.2. Periodic Boundary Conditions. In the present study, the term “periodic boundary conditions” refers to a system with a potential energy U_{pot} of the form

$$U_{\text{pot}} = \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \sum_{\mathbf{n}}' v(\mathbf{r}_{ij} + a\mathbf{n}, \omega_i, \omega_j) \quad (2)$$

where $\mathbf{n} = (n_x, n_y, n_z)$ is a vector that runs over all lattice points in the particular (unit length) lattice and a denotes the side length of the unit cell. Furthermore, the primed sum indicates that the term with $i = j$ for $\mathbf{n} = \mathbf{0}$ should be excluded, and $v(\mathbf{r}_{ij}, \omega_i, \omega_j)$ denotes the intermolecular potential between particles i and j , depending in general on their separation \mathbf{r}_{ij} and orientations ω_i and ω_j . In practice, $v(\mathbf{r}_{ij}, \omega_i, \omega_j)$ is usually long-range in the sense that it decays no faster than r^{-3} , the two most important examples being the Coulomb and dipole–dipole potentials.

Since the sum in eq 2 is slowly (and conditionally) convergent, more elaborate methods to evaluate the potential energy in a PBC system need to be used in practice. The by far most popular technique to achieve a fast convergence of the potential energy is the technique originally due to Ewald⁶ and different mesh-based variants^{7–10} thereof. Within the Ewald-based methods, the short-range ($\mathbf{n} = \mathbf{0}$) part of U_{pot} is screened through the addition of a Gaussian charge (dipole) cloud and is thereafter summed within a, usually spherical, cutoff after considering the nearest image convention.¹ The long-range ($\mathbf{n} \neq \mathbf{0}$) part of the potential energy is summed up in Fourier space, leading to a quickly (and absolutely) convergent sum.

2.3. Simulation Cells with Noncubic Geometries. Although cubic simulation cells are used for the majority of simulation studies, the use of alternative simulation cell geometries

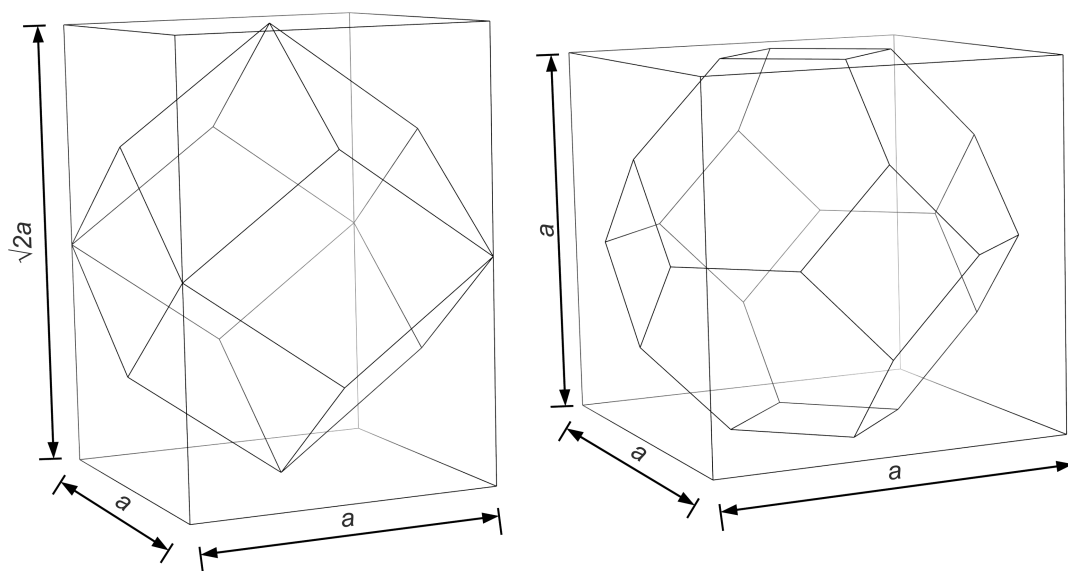


Figure 2. The rhombic dodecahedron (left) and truncated octahedron (right) inscribed in their BCT and BCC unit cells, respectively.

has started to become increasingly popular. In total, five classes of geometrical bodies are translationally space-filling and can thus be used for simulating a periodic system;³¹ however, due to their relatively “sphere-like” symmetry, the two most useful alternatives to the cube, at least for the simulation of bulk systems, are the rhombic dodecahedron (RD) and truncated octahedron (TO). While the cube, of course, packs in a simple cubic (SC) lattice structure, the natural choice for the lattice structures of the RD and TO are face-centered cubic (FCC) and body-centered cubic (BCC), respectively.³¹ However, Smith and Fincham³⁶ showed that the use of a body-centered tetragonal (BCT) lattice structure for the RD, with one side of the unit cell elongated by a factor $\sqrt{2}$ compared to the other two, greatly facilitates the implementation of Ewald summation for this geometry, by simply excluding k -space terms of certain parity. Thus, we will use the BCC and BCT lattice structures as the basis of our analysis. In Figure 2, the RD and the TO are shown, inscribed in their respective unit cells.

2.4. Solvation in Dielectric Media. In the following subsections, we will treat relevant parts of the theory of solvation and fluctuations in dielectric media. First, we will review the theory for the solvation of a polarizable dipole in a dielectric medium (section 2.4.1). Subsequently, in section 2.4.2, we will treat electrostatic fluctuations and solvation in isotropic dielectric media. Finally, in section 2.4.3, we will use tools from the two preceding parts to develop a heuristic model for the solvation of a dielectric subvolume in a PBC system.

Generally, the collective electrostatic fluctuations will be quantified through the spherical multipole moments Q_{lm} , defined through

$$Q_{lm} \equiv \int_V \rho(\mathbf{r}) r^l C_{lm}(\Omega) d\mathbf{r} \quad (3)$$

where $\rho(\mathbf{r})$ denotes the charge density in a point $\mathbf{r} = (r, \Omega) = (r, \varphi, \theta) \in V$ and $C_{lm}(\Omega)$ represents Racah’s unnormalized spherical harmonics. The index l denotes the order of the multipole, whereas m describes its orientation in an external coordinate frame. Just as for the spherical harmonics, m takes on all integer values between $-l$ and $+l$. However,

the $-m$ and $+m$ components are related according to

$$Q_{l-m} = (-1)^m Q_{lm}^* \quad (4)$$

where $*$ denotes complex conjugation; thus, Q_{lm} and Q_{l-m} are not independent degrees of freedom. Instead, we will adopt the approach taken previously^{37,38} and treat separately the real and imaginary parts of Q_{lm} , denoted respectively by superscripts R and I, for $m \geq 0$. Since Q_{l0} is real, these $l+1$ real and l imaginary multipole components form $2l+1$ linearly independent *fluctuation modes*. Furthermore, we will use the bracketed superscripts (R) and (I) to denote quantities that are somehow related to the real and imaginary parts of Q_{lm} , although not themselves complex quantities.

2.4.1. A Polarizable Dipole in a Dielectric Medium. The solvation energy U_{solv} of a polarizable point dipole of magnitude μ , radius R , and polarizability α embedded in a dielectric medium of dielectric permittivity ϵ is given by³⁹

$$U_{\text{solv}} = -\frac{1}{2} \frac{g\mu^2}{1-g\alpha} \quad (5)$$

where

$$g = \frac{1}{R^3} \frac{2(\epsilon-1)}{2\epsilon+1} \quad (6)$$

quantifies the reaction field, parallel to the dipole, coming from the surrounding dielectric medium. A physical interpretation of the expression for U_{solv} is facilitated by expanding eq 5 in a geometric series, i.e.

$$U_{\text{solv}} = -\frac{g\mu^2}{2} \sum_{n=0}^{\infty} (g\alpha)^n \quad (7)$$

From this expression, we can identify the prefactor $-g\mu^2/2$ as the solvation energy of a *permanent* dipole immersed in a dielectric medium, whereas the factor $\sum_n (g\alpha)^n$ takes into account the increase of the solvation energy due to the additional polarization of the particle by the reaction field. The infinite sum is due to the incremental nature of this process; the reaction field increases the total dipole moment of the particle, which in

turn polarizes the dielectric to yield a larger reaction field, etc. In section 2.4.3, we will show how this partitioning of the solvation energy can be mapped onto the solvation of a dielectric subvolume exhibited to PBCs.

2.4.2. Electrostatic Fluctuations in a Dielectric Medium. The (unnormalized) probability distribution $P_{\text{vac}}(Q_{lm}^{\text{X}})$ of the 2^l -pole moment of a spherical dielectric volume with radius R and dielectric permittivity ε in a vacuum is given by the Gaussian function^{37,40}

$$P_{\text{vac}}(Q_{lm}^{\text{X}}) = \exp[-\beta U_{\text{vac}}^{\dagger}] \quad (8)$$

where U_{vac}^{\dagger} given by

$$U_{\text{vac}}^{\dagger} = \frac{(2 - \delta_{m0})[l(\varepsilon + 1) + 1]}{2l(\varepsilon - 1)} \frac{(Q_{lm}^{\text{X}})^2}{R^{2l+1}} \quad (9)$$

denotes the (free) energy cost for creating an instantaneous multipole moment Q_{lm}^{X} in the dielectric volume, $\beta = (k_{\text{B}}T)^{-1}$ is the inverse thermal energy, and $\text{X} \in \{\text{R}, \text{I}\}$. If the dielectric volume is immersed in an infinite dielectric medium with the same value of ε as the sphere itself, U^{\dagger} is decreased due to the depolarizing reaction field from the surrounding medium, changing the probability distribution to^{37,40}

$$P_{\text{diel}}(Q_{lm}^{\text{X}}) = \exp[-\beta U_{\text{diel}}^{\dagger}] \quad (10)$$

where

$$\begin{aligned} U_{\text{diel}}^{\dagger} &= \frac{(2 - \delta_{m0})(2l + 1)^2 \varepsilon}{2(\varepsilon - 1)l[(l + 1)\varepsilon + l]} \frac{(Q_{lm}^{\text{X}})^2}{R^{2l+1}} \\ &\approx (2 - \delta_{m0}) \frac{(2l + 1)^2}{2l(l + 1)\varepsilon} \frac{(Q_{lm}^{\text{X}})^2}{R^{2l+1}} \equiv c_{lm} \times (Q_{lm}^{\text{X}})^2 \end{aligned} \quad (11)$$

and the second equality is accurate for not too small values of ε . The energy expression in eq 9 is roughly independent of ε for high-dielectric media and thus not numerically useful for determining ε from a computer simulation. In contrast, the right-hand-side of eq 11 shows that $U_{\text{diel}}^{\dagger} \sim \varepsilon^{-1}$ for large and intermediate values of ε , and thus determining the width of P_{diel} in a computer simulation can be used to determine the dielectric permittivity of the system under study. More specifically, eq 10 can be transformed into a formula for the mean-square quantity $\langle (Q_{lm}^{\text{X}})^2 \rangle$ by noting that the Gaussian form implies that $\langle (Q_{lm}^{\text{X}})^2 \rangle = (2\beta c_{lm})^{-1}$. After some rearrangements, still using the simplified form for high and intermediate ε , we get

$$\varepsilon \approx (2 - \delta_{m0}) \frac{(2l + 1)^2}{l(l + 1)} \frac{\beta \langle (Q_{lm}^{\text{X}})^2 \rangle}{R^{2l+1}} \quad (12)$$

The $l = 1$ case of eq 12 applied to the *total* dipole moment has been widely used to determine ε from computer simulations of fluids, although care needs to be taken to use a form of the formula proper for the particular boundary conditions being used.^{41,42} In an infinite, isotropic dielectric medium, ε is by definition independent of l , m , and R , but for a finite and/or molecular system, this does not necessarily hold. In particular, as we will show below, ε is not independent of l and m for a system exposed to PBCs.

In addition to the dielectric permittivity, the above formulas can be used to obtain the free energy change $\Delta A_{\text{vac} \rightarrow \text{diel}}$ of

bringing the dielectric sphere from vacuum into its own medium. To this end, we will employ the standard relationship⁴³

$$\beta \Delta A_{\text{vac} \rightarrow \text{diel}} = -\ln \frac{Z_{\text{diel}}}{Z_{\text{vac}}} = -\ln \frac{\int_{-\infty}^{\infty} P_{\text{diel}}(Q_{lm}^{\text{X}}) dQ_{lm}^{\text{X}}}{\int_{-\infty}^{\infty} P_{\text{vac}}(Q_{lm}^{\text{X}}) dQ_{lm}^{\text{X}}} \quad (13)$$

where Z_{diel} and Z_{vac} denote the configuration integrals in the solvated and nonsolvated states, respectively. Inserting eqs 8–11 and carrying out the integrations gives

$$\begin{aligned} \beta \Delta A_{\text{vac} \rightarrow \text{diel}} &= \frac{1}{2} \ln \left[\frac{(2l + 1)^2 \varepsilon}{[(l + 1)\varepsilon + l][l(\varepsilon + 1) + 1]} \right] \\ &\approx \frac{1}{2} \ln \left[\frac{(2l + 1)^2}{l(l + 1)\varepsilon} \right] \end{aligned} \quad (14)$$

Thus, $\Delta A_{\text{vac} \rightarrow \text{diel}}$ is (i) always negative, (ii) independent of m and R , and (iii) only weakly dependent on l , a dependence that disappears quickly in the limit $l \rightarrow \infty$. Finally, we note that $\Delta A_{\text{vac} \rightarrow \text{diel}}$ diverges logarithmically as $\varepsilon \rightarrow \infty$ for all l .

2.4.3. Electrostatic Fluctuations in a System Subjected to PBCs. We will now propose a mapping of the energy expression in eq 7 for a polarizable dipole in a dielectric medium onto the solvation of a dielectric subvolume in a system exposed to PBCs. As a first assumption, we will describe the energy of creating an instantaneous multipole moment Q_{lm}^{X} in the spherical volume, *excluding* the anisotropic part of the solvation, by the same expression as in an infinite dielectric medium. Using eq 11 and eq 7, we thus make the assignment

$$-\frac{g\mu^2}{2} \sim U_{\text{diel}}^{\dagger} = (2 - \delta_{m0}) \frac{(2l + 1)^2}{2l(l + 1)\varepsilon} \frac{(Q_{lm}^{\text{X}})^2}{R^{2l+1}} \quad (15)$$

Obviously, $U_{\text{diel}}^{\dagger}$ is qualitatively different from the prefactor $-g\mu^2/2$ of eq 7; most importantly, it has a positive rather than a negative sign, since it also includes the energetic cost of creating the multipole moment in the dielectric medium, whereas the energy in eq 7 is valid for a permanent dipole, i.e., excluding the self-energy of the charge distribution.

In addition to the isotropic solvation, the instantaneous multipole moment induces a generalized reaction field coming from its own replicas in all of the surrounding boxes, which in turn polarizes the dielectric volume. This behavior is fully analogous to the polarization of a polarizable dipole by its own reaction field; however, in the case of PBCs the reaction field is not proportional to the factor g of eq 6 but rather to the lattice interaction tensor S_{lm}^{X} quantifying the interaction between the multipole component Q_{lm}^{X} and all its replicas in the lattice. The use of eqs 44 and 47 of ref 38 leads us to the following definition of S_{lm}^{X} :

$$\begin{aligned} S_{lm}^{\text{X}} &\equiv (-1)^{l+\delta_{\text{Xl}}} f(l, l, m, m) \sum_{\mathbf{n} \neq 0} \frac{C_{2l, 2m}^{\text{R}}(\Omega)}{|\mathbf{a}\mathbf{n}|^{2l+1}} \\ &+ (1 - \delta_{m0}) (-1)^{l+m} f(l, l, m, -m) \sum_{\mathbf{n} \neq 0} \frac{C_{2l, 0}^{\text{R}}(\Omega)}{|\mathbf{a}\mathbf{n}|^{2l+1}} \end{aligned} \quad (16)$$

where a is the side length of the unit cell, δ is the Kronecker delta,

and the function f is defined by

$$f(l_1, l_2, m_1, m_2) \equiv \left[\frac{(2(l_1 + l_2))!}{(2l_1)!(2l_2)!} \right]^{1/2} \sqrt{2(l_1 + l_2) + 1} \begin{pmatrix} l_1 & l_2 & l_1 + l_2 \\ m_1 & m_2 & -m_1 - m_2 \end{pmatrix} \quad (17)$$

with (\dots) representing the Wigner $3j$ symbol.⁴⁴ The two terms for $m > 0$ come from the interaction between $Q_{lm}^{\mathbf{x}}$ in the central unit cell and (i) $Q_{lm}^{\mathbf{x}}$ and (ii) $Q_{l-m}^{\mathbf{x}}$ in the surrounding cells. We furthermore note that (see Appendix A)

$$\sum_{m, \mathbf{x}} S_{lm}^{(\mathbf{x})} = 0, \quad \forall l \geq 1 \quad (18)$$

where the sum runs over all multipole components with a given l . Thus, the (unweighted) mean value of the lattice interaction for any $l \geq 1$ is zero for all multipoles and lattices. It should finally be clarified that, whereas g always yields an attractive coupling between the polarizable dipole and the reaction field, $S_{lm}^{(\mathbf{x})}$ can represent attractive as well as repulsive couplings, depending on the symmetry properties of each multipole. Finally, we make the assumption that the polarizability α in eq 7 can be mapped according to $\alpha \sim k_{\text{self}} R^{2l+1}$, where k_{self} is a positive constant related to the magnitude of the anisotropic solvation.

On the basis of the above discussion, we suggest that the (free) energy for creating an instantaneous multipole moment in a spherical subvolume of a dielectric exhibited to PBCs is given by

$$U_{\text{PBC}}^{\dagger} = (2 - \delta_{m0}) \frac{(2l + 1)^2}{2l(l + 1)\epsilon} \frac{(Q_{lm}^{\mathbf{x}})^2}{R^{2l+1}} \frac{1}{1 - k_{\text{self}} R^{2l+1} S_{lm}^{(\mathbf{x})}} \quad (19)$$

In accordance with eq 10, we also form the corresponding probability distribution $P_{\text{PBC}}(Q_{lm}^{\mathbf{x}})$:

$$P_{\text{PBC}}(Q_{lm}^{\mathbf{x}}) = \exp[-\beta U_{\text{PBC}}^{\dagger}] \quad (20)$$

We note that, just like in the case of isotropic dielectric solvation, P_{PBC} is Gaussian (as long as $k_{\text{self}} R^{2l+1} S_{lm}^{(\mathbf{x})} < 1$), but with the important difference that its exponent is now m -dependent through the dependence on $S_{lm}^{(\mathbf{x})}$. The Gaussian form with respect to $Q_{lm}^{\mathbf{x}}$ implies that the mean-square multipole moment $\langle (Q_{lm}^{\mathbf{x}})^2 \rangle_{\text{PBC}}$ can be expressed as

$$\langle (Q_{lm}^{\mathbf{x}})^2 \rangle_{\text{PBC}} = \frac{R^{2l+1}}{(2 - \delta_{m0})\beta} \frac{l(l + 1)\epsilon}{(2l + 1)^2} [1 - k_{\text{self}} R^{2l+1} S_{lm}^{(\mathbf{x})}] \quad (21)$$

In analogy with eq 12, we now define the *apparent* dielectric permittivity $\epsilon_{\text{PBC},lm}^{(\mathbf{x})}$ as

$$\epsilon_{\text{PBC},lm}^{(\mathbf{x})} \equiv (2 - \delta_{m0}) \frac{(2l + 1)^2}{l(l + 1)} \frac{\beta \langle (Q_{lm}^{\mathbf{x}})^2 \rangle_{\text{PBC}}}{R^{2l+1}} \quad (22)$$

which from the above reasoning now becomes dependent on l and m due to the anisotropic polarization induced by the PBCs. Finally, inserting eq 21 into eq 22 gives the relation

$$\epsilon_{\text{PBC},lm}^{(\mathbf{x})} = \epsilon_{\text{diel}} [1 - k_{\text{self}} R^{2l+1} S_{lm}^{(\mathbf{x})}] \quad (23)$$

where we have added the subscript “diel” to ϵ , to stress that it represents the *true* (m -independent) dielectric permittivity that the fluid would have if it behaved as an isotropic dielectric

medium. For a molecular system, $\epsilon_{\text{PBC},lm}^{(\mathbf{x})}$ can be obtained by sampling $\langle (Q_{lm}^{\mathbf{x}})^2 \rangle_{\text{PBC}}$ in a computer simulation. By plotting $\epsilon_{\text{PBC},lm}^{(\mathbf{x})}$ as a function of $S_{lm}^{(\mathbf{x})}$, the two constants ϵ_{diel} and k_{self} appearing in eq 23 can be determined from the intercept and slope of a linear fit to the data points.

Just as in the case of dielectric solvation, we may use the analogy of eq 13 to define the free energy change $\Delta A_{\text{vac} \rightarrow \text{PBC}}$ of bringing a dielectric sphere from a vacuum into a system under PBCs. Using eqs 8, 9, and 19–20 gives, after performing the integrations,

$$\beta \Delta A_{\text{vac} \rightarrow \text{PBC}} \approx \frac{1}{2} \ln \left[\frac{(2l + 1)^2}{l(l + 1)\epsilon_{\text{diel}}(1 - k_{\text{self}} R^{2l+1} S_{lm}^{(\mathbf{x})})} \right] \quad (24)$$

Since $\Delta A_{\text{vac} \rightarrow \text{PBC}}$ describes the long-range part of the electrostatic free energy of the simulated system, it should ideally not differ too much from $\Delta A_{\text{vac} \rightarrow \text{diel}}$, and therefore a comparison between these two quantities may be a good way of assessing the accuracy of the particular boundary conditions being used. In particular, for $S_{lm}^{(\mathbf{x})} = 0$ or $k_{\text{self}} = 0$, $\Delta A_{\text{vac} \rightarrow \text{PBC}}$ reduces to eq 14 for an isotropic dielectric medium.

3. COMPUTATIONAL DETAILS

The molecular model system is composed by particles possessing a pairwise additive interparticle potential $v(r_{ij}, \omega_i, \omega_j)$, composed of a dipolar and a Lennard-Jones (LJ) part according to

$$v(r_{ij}, \omega_i, \omega_j) = v_{\text{dip}}(r_{ij}, \omega_i, \omega_j) + v_{\text{LJ}}(r_{ij}) \quad (25)$$

where

$$v_{\text{dip}}(r_{ij}, \omega_i, \omega_j) = \left[\frac{\boldsymbol{\mu}_i \cdot \boldsymbol{\mu}_j}{r_{ij}^3} - \frac{3(\boldsymbol{\mu}_i \cdot \mathbf{r}_{ij})(\boldsymbol{\mu}_j \cdot \mathbf{r}_{ij})}{r_{ij}^5} \right] \quad (26)$$

and

$$v_{\text{LJ}}(r_{ij}) = 4\epsilon_{\text{LJ}} \left[\left(\frac{\sigma_{\text{LJ}}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{\text{LJ}}}{r_{ij}} \right)^6 \right] \quad (27)$$

In the above equations, $\boldsymbol{\mu}_i$ represents the dipole of particle i , \mathbf{r}_{ij} is the vector pointing from particle i to particle j , $r_{ij} = |\mathbf{r}_{ij}|$, and ϵ_{LJ} and σ_{LJ} are the LJ parameters. Two different values of the molecular dipole moment $\mu = |\boldsymbol{\mu}|$ were employed: $\mu = 0.45$ atomic units ($0.23813e \text{ \AA}$, $\mu^* \equiv \mu / (4\pi\epsilon_0\epsilon_{\text{LJ}}\sigma_{\text{LJ}}^3)^{1/2} = 1.290$) and $\mu = 0.65$ atomic units ($0.34397e \text{ \AA}$, $\mu^* = 1.863$). The LJ parameters were set to $\sigma_{\text{LJ}} = 2.8863 \text{ \AA}$ and $\epsilon_{\text{LJ}} = 1.97023 \text{ kJ mol}^{-1}$.

The thermodynamic properties of the model system were determined by performing MD simulations in the canonical (constant N, V, T) ensemble, using $N = 1000$ particles in a cell of volume $V = 2.601 \times 10^4 \text{ \AA}^3$ for all three simulation cell geometries. The temperature was kept constant at $T = 315.78 \text{ K}$ ($T^* \equiv k_{\text{B}}T/\epsilon_{\text{LJ}} = 1.333$). Toroidal boundaries for the noncubic simulation cells were applied according to the procedures devised by Smith,³² whereas Ewald summation with tin foil boundaries were implemented using the formulas due to Smith and Fincham.³⁶ A spherical cutoff in real space of $r_{\text{cut}} = 14 \text{ \AA}$ was used in conjunction with the Ewald screening parameter $\alpha = 3.2/r_{\text{cut}}$. The cutoff n_{cut} in reciprocal space was set to 7, 10, and 9 for the cube, RD, and TO geometries, respectively, to yield a constant relative error in the k -space energy of $\sim 10^{-5}$. For all

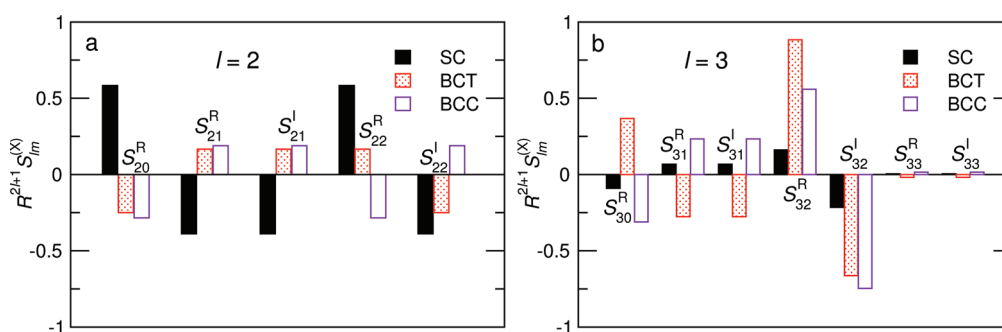


Figure 3. Values of the reduced interaction tensor $R^{2l+1}S_{lm}^{(X)}$ for (a) $l = 2$ and (b) $l = 3$ relevant for the three different simulation cell geometries. The values were obtained from eq 16 using a spherical cutoff of $n_{\max} = 50$.

simulations, the integrated MC/MD/Brownian dynamics simulation package Molsim⁴⁵ was used. For further details about the simulation parameters, the reader is referred to our previous study.²⁸

The multipole moments Q_{lm} , $1 \leq l \leq 4$, of a sphere with radius R were evaluated after every 100th time step. The contribution $Q_{lm,i}$ from a molecular dipole $Q_{1m,i}$ located at $\mathbf{r} = (r, \Omega)$ to the total multipole moment $Q_{lm} = \sum_i Q_{lm,i}$ was calculated according to²⁸

$$Q_{lm,i} = \sum_{m'=-1}^1 (-1)^{l+m} [l(2l-1)(2l+1)]^{1/2} \times \begin{pmatrix} l-1 & l & 1 \\ m+m' & -m & -m' \end{pmatrix} \times Q_{1,-m'} r^{l-1} C_{l-1,m+m'}(\Omega) \quad (28)$$

where all terms containing C_{lm} with $|m| > l$ should be excluded. For each sampled configuration, Q_{lm} was calculated with each particle used as the origin, giving in total N sampled values of Q_{lm} per configuration. In addition, reference values of ϵ_{diel} for various R values were calculated using eq 12 from a simulation in a cubic simulation cell and $N = 10^5$ particles, i.e., 100 times as large as the primary systems. This large (compared to R) system size was used in order to ensure that the values of ϵ_{diel} thus obtained are unaffected by the boundary. This system is further described in conjunction with our previous study.²⁸

4. RESULTS AND DISCUSSION

In Figure 3, numerically calculated values of the reduced interaction tensor $R^{2l+1}S_{lm}^{(X)}$ for $l = 2$ and 3 are given for the SC, BCT, and BCC lattices. We note that $R^{2l+1}S_{lm}^{(X)}$ is highly dependent on m , taking on both positive, corresponding to repulsive net interaction energies, and negative, corresponding to attractive net interactions, values. We also note that there is no obvious correlation between either the sign or the magnitude of $R^{2l+1}S_{lm}^{(X)}$ obtained from the three different lattices. For $l = 2$, the SC lattice yields significantly ($\approx 100\%$) higher absolute values of $R^{2l+1}S_{lm}^{(X)}$ than the other two lattices; however, for $l = 3$, the situation is the opposite. This means that the magnitude (and sign) of the coupling of a certain multipole depends strongly on its symmetry in relation to the symmetry of the particular lattice where it resides. We furthermore note (results not shown) that $S_{lm}^{(X)} = 0$ for all m and all lattices, meaning that the total dipole–dipole interaction is zero. This is a well-known fact for all

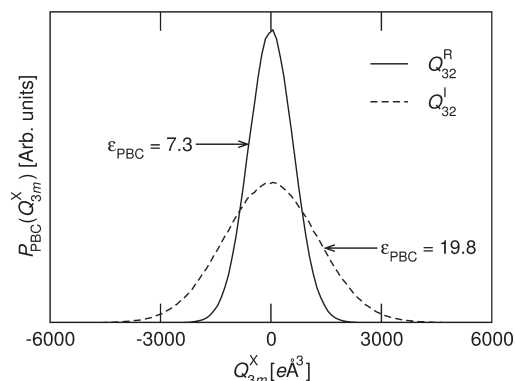


Figure 4. Probability distribution $P_{\text{PBC}}(Q_{lm}^X)$ of two octupole components obtained from a simulation with $\mu = 0.45$ in RD geometry. The values of the corresponding reduced interaction tensors are $R^{2l+1}S_{32}^{(R)} = 0.88$ and $R^{2l+1}S_{32}^{(I)} = -0.66$.

cubic lattices, as long as the lattice sum is carried out in spherical shells.^{24–27}

In Figure 4, the probability distribution $P_{\text{PBC}}(Q_{lm}^X)$ for two different octupole components, obtained from a simulation of the $\mu = 0.45$ system in RD geometry, is shown. Clearly, there is a significant difference between the widths of the two probability distributions, although both distributions follow the predicted Gaussian form very well. The width of $P_{\text{PBC}}(Q_{lm}^X)$ should be compared to the corresponding values of $S_{lm}^{(X)}$ given in Figure 3. Obviously, the narrower one of the two probability distributions corresponds to a repulsive value of $S_{lm}^{(X)}$ ($R^{2l+1}S_{32}^{(R)} = 0.88$), whereas the wider distribution corresponds to an attractive net interaction ($R^{2l+1}S_{32}^{(I)} = -0.66$). Thus, there is at least qualitative reason in our assumption that the width of $P_{\text{PBC}}(Q_{lm}^X)$, and thus the magnitude of ϵ_{PBC} , can be described by the lattice interaction tensor $S_{lm}^{(X)}$.

In order to quantitatively assess the dependence of $P_{\text{PBC}}(Q_{lm}^X)$ on $S_{lm}^{(X)}$, we evaluated the former quantity in terms of the apparent (m -dependent) dielectric permittivity $\epsilon_{\text{PBC},lm}^{(X)}$ defined through eq 22. In Figure 5, plots of $\epsilon_{\text{PBC},lm}^{(X)}$ versus $R^{2l+1}S_{lm}^{(X)}$ obtained for both dipole strengths and all three simulation cell geometries are presented. Clearly, the proposed linear relationship between $\epsilon_{\text{PBC},lm}^{(X)}$ and $R^{2l+1}S_{lm}^{(X)}$ is very well reproduced by the simulation data in all cases. Furthermore, the results obtained from the $\mu = 0.45$ systems exhibit slopes that are essentially independent of geometry and l . The slopes of the $\mu = 0.65$ data show a larger variation, although no systematic dependence on

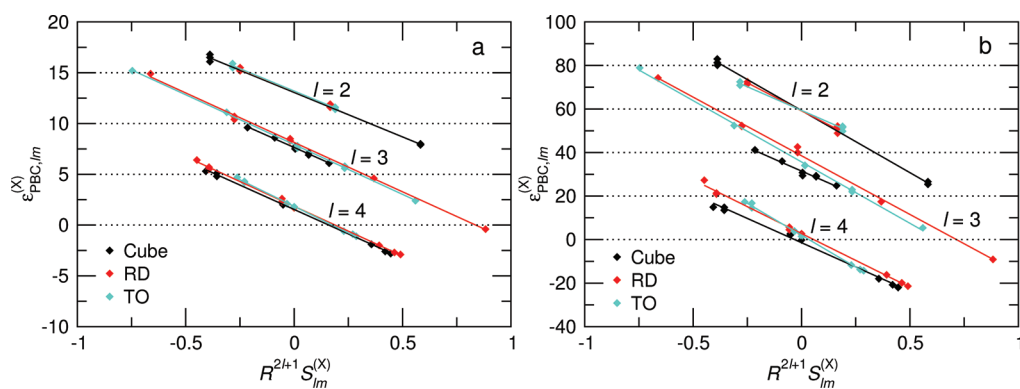


Figure 5. Apparent dielectric permittivity $\epsilon_{\text{PBC},lm}^{(X)}$ obtained from simulations of dipoles with (a) $\mu = 0.45$ and (b) $\mu = 0.65$ versus $R^{2l+1}S_{lm}^{(X)}$ for the corresponding lattice types. The results for $l = 3$ ($l = 4$) have been shifted vertically by 5 (10) units for $\mu = 0.45$ and 20 (40) units for $\mu = 0.65$ to enhance readability.

Table 1. Fitted Values (Figure 5 and eq 23) of k_{self} and ϵ_{diel} for Dipoles with $\mu = 0.45$ and 0.65 and Values of ϵ_{diel} Obtained (eq 12) from a Simulation in Cubic Geometry with $N = 10^5$ Particles

μ	geometry	R [Å]	l	k_{self}	$\epsilon_{\text{diel}}^{(\text{fit})}$	$\epsilon_{\text{diel}}^{(\text{ref})}$
0.45	cube	14.80	2	0.67	13.1	13.4
			3	0.75	12.6	12.5
			4	0.83	11.5	11.7
	RD	16.60	2	0.66	13.2	13.7
			3	0.74	13.1	12.9
			4	0.83	11.8	12.1
TO	16.16	2	0.67	13.2	13.6	
		3	0.76	12.9	12.8	
		4	0.89	11.8	12.0	
0.65	cube	14.80	2	0.96	59	67
			3	0.85	52	51
			4	1.16	39	39
	RD	16.60	2	0.87	59	73
			3	0.92	59	56
			4	1.15	43	44
TO	16.16	2	0.73	59	72	
		3	1.01	56	55	
		4	1.40	42	43	

geometry and l is apparent. We also note the perhaps somewhat nonintuitive fact that the magnitude of the self-interaction (quantified through $R^{2l+1}S_{lm}^{(X)}$) does not decay with increasing l , at least not for $l \leq 4$. Furthermore, the self-interaction magnitudes do not show any clear trend between the different cell geometries. As an example, we note that the cubic geometry exhibits the largest quadrupole ($l = 2$) self-interactions, whereas the octupole ($l = 3$) self-interaction has its largest magnitude in the RD and TO geometries.

Because of the good linearity of the data, eq 23 can be used to obtain values of ϵ_{diel} and k_{self} from the intercept and slope, respectively, of the data in Figure 5. In Table 1, fitted values of ϵ_{diel} and k_{self} are given for both dipole strengths and all three cell geometries, together with values of ϵ_{diel} independently calculated from a simulation with $N = 10^5$ using eq 12 and the same values of the sampling radius R . From this data, we note that

1. The fitted values of ϵ_{diel} are close (within 5%) to the ones calculated from eq 12 for all fittings except those with

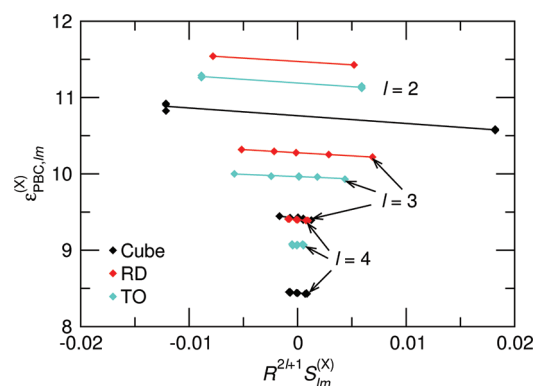


Figure 6. Data corresponding to Figure 5a but using sampling radii R half as large. Note the different scale on both axes and that the data have *not* been shifted in the y direction.

$\mu = 0.65$ and $l = 2$, where the fittings generally yield too low values of ϵ_{diel} .

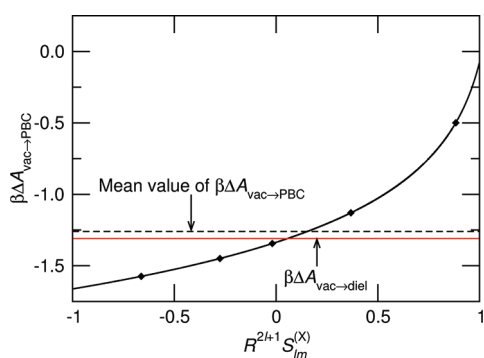
2. The fitted values of k_{self} are slightly larger and exhibit larger variations for $\mu = 0.65$ than for $\mu = 0.45$. Observation 1 shows that our assumption that the interaction between noncorrelated multipoles (i.e. excluding the self-interaction) can be described using formulas for an isotropic dielectric medium (eq 15) is reasonable, perhaps with the exception for the $l = 2$ and $\mu = 0.65$ case. Observation 2 indicates that there may be a slight variation of k_{self} with ϵ_{diel} (or μ), although the source of this variation is not clear. The larger variation in k_{self} for $\mu = 0.65$ we attribute to the larger statistical noise present in the more strongly coupled system.

We furthermore note that ϵ_{diel} decreases with increasing l , in line with what we have observed before.^{28,46} The apparent geometry dependence of ϵ_{diel} is not due directly to the geometry but rather to the slightly different radii of the inscribed spheres in the three simulation cells; this behavior is also consistent with our previous observation^{28,46} that ϵ_{diel} increases with increasing sampling sphere radius R for a given l .

In Figure 6, the data corresponding to Figure 5a, but obtained using sampling radii half as large, are presented. In this case, it is obvious that the magnitude of the self-interaction quickly becomes less significant for increasing l , due to its $R^{-(2l+1)}$ dependence. The linear fits are however still satisfactory, even

Table 2. Data As in Table 1 but Obtained Using Sampling Radii Half As Large and Only for $\mu = 0.45$

geometry	R [Å]	l	k_{self}	$\epsilon_{\text{diel}}^{(\text{fit})}$	$\epsilon_{\text{diel}}^{(\text{ref})}$
cube	7.40	2	0.94	10.8	10.8
		3	1.84	9.4	9.4
		4	1.34	8.4	8.4
RD	8.80	2	0.77	11.5	11.5
		3	0.79	10.3	10.3
		4	1.33	9.4	9.4
TO	8.08	2	0.86	11.2	11.2
		3	0.59	10.0	10.0
		4	0.06	9.1	9.1

**Figure 7.** Solvation free energy $\Delta A_{\text{vac} \rightarrow \text{PBC}}$ for $l = 3$ (black solid line) obtained from a $\mu = 0.65$ system in RD geometry using eq 24 and values of k_{self} and $\epsilon_{\text{diel}}^{(\text{fit})}$ from Table 1. The red solid line gives $\Delta A_{\text{vac} \rightarrow \text{diel}}$ obtained from eq 14 using $\epsilon_{\text{diel}}^{(\text{sim})}$ from Table 1, and the black dashed line gives the mean value of $\Delta A_{\text{vac} \rightarrow \text{PBC}}$, averaged over all seven octupole components (black symbols, two doubly degenerate values).

though the very small variation in $\epsilon_{\text{PBC},lm}^{(X)}$ over the range of $R^{2l+1}S_{lm}^{(X)}$ values leads to larger statistical errors in the fittings, especially for $l = 3$ and 4. The apparent shift in the y direction between curves obtained using different geometries is merely due to the values used for the sampling radius R being geometry dependent, leading to different values of the intrinsic dielectric permittivity ϵ_{diel} . In fact, the same shift is present in Figure 5, although it is not visible due to the much wider range of the ordinate axis. According to our theoretical assumptions, the value of k_{self} should be independent of R , meaning that the slope of the lines in Figures 5a and 6 should be identical. In Table 2, the values of ϵ_{diel} and k_{self} obtained from the smaller sampling radii are presented. Although the variation in k_{self} is larger than in Table 1 due to the larger statistical noise, our assumption for k_{self} is not obviously contradicted. Using the smaller sampling radii for the $\mu = 0.65$ system (data not shown), however, seems to yield somewhat larger values of k_{self} than in Table 1, although the statistical significance of these values can be questioned.

Figure 7 gives the free energy $\Delta A_{\text{vac} \rightarrow \text{PBC}}$ for $l = 3$, $\mu = 0.65$, and RD geometry calculated using eq 24. Clearly, the anisotropy in ϵ_{PBC} discussed in the previous paragraphs also corresponds to a large anisotropy in the solvation free energy of the subvolume. Quantitatively, $\beta\Delta A_{\text{vac} \rightarrow \text{PBC}}$ varies between -1.6 and -0.5 , compared to the dielectric value $\beta\Delta A_{\text{vac} \rightarrow \text{diel}} \approx -1.30$ (eq 14, red solid line in Figure 7). However, the average of $\beta\Delta A_{\text{vac} \rightarrow \text{PBC}}$ over all seven octupole components is $\beta\Delta A_{\text{vac} \rightarrow \text{PBC}}^{(\text{avg})} \approx -1.26$ (black dashed line in Figure 7), i.e., very close to the dielectric

value. Thus, the *total* solvation free energy (at least on the octupolar level) of the simulation box is very close to that of an isotropic system, even though it is distributed in a highly anisotropic way. A possible key to understanding this behavior is to be found in eq 18, namely, that the (unweighted) interaction tensors for any given $l \geq 1$ cancel out when summed over all multipole components. Thus, the suppression of some fluctuation modes is exactly compensated by the enhancement of others, leading to a reasonable “mean value” of the energy. This behavior is reproduced for all multipole components and geometries (results not shown), in the sense that $\beta\Delta A_{\text{vac} \rightarrow \text{PBC}}^{(\text{avg})}$ and $\beta\Delta A_{\text{vac} \rightarrow \text{diel}}$ are always within 10% of each other.

5. CONCLUSIONS

In the present study, we have presented a quantitative analysis of the periodicity effects induced in a dipolar system by the use of PBCs. Using classical electrostatics and statistical thermodynamics, we developed a heuristic model relating the apparent, anisotropic dielectric permittivity $\epsilon_{\text{PBC},lm}^{(X)}$ to the reduced lattice interaction tensors $S_{lm}^{(X)}$. The theory exhibits excellent agreement with results from MD simulations of Stockmayer fluids with two different dipole strengths and three different simulation cell geometries. Although the anisotropy in the electrostatic fluctuations is independent of l on the length scale of the simulation box, it is shown that the “range” of the boundary effects (i.e., the minimum value of R needed to induce significant boundary effects) decreases strongly with increasing l . Furthermore, it was shown that the large ($\sim 200\%$) anisotropy in the solvation free energy on the length scale of the simulation box disappears when averaged over all fluctuation modes, leading to the total solvation free energy being practically identical to the value predicted for an isotropic system.

Even though the simulation part of our study is based on a Stockmayer model system, we argue that our use of a dielectric continuum model as the theoretical basis means that the effects are fully transferable to any polar system which may be described as a dielectric medium, in particular the many popular water models used in molecular simulations. We also expect that any structural property, i.e., not only the dielectric permittivity, evaluated on the length-scale of the full simulation box is equally affected by the boundary effects.

One of the most important observations from this study is that the total solvation free energy is, in spite of the large anisotropy of the individual contributions, very close to the correct, isotropic value. This observation is indeed closely analogous to the corresponding averaging in the anisotropy of the radial distribution function for a Lennard-Jones fluid under PBCs observed by Pratt and Haan.¹³ We argue that this property explains the success of Ewald summation and related techniques, at least when it comes to evaluating energies and relatively short-range structural properties. Nevertheless, as we have also shown previously,²⁸ one should use caution when evaluating structural properties on length scales larger than half the length of the simulation box.

Another relevant question is whether there is any rationale behind using a noncubic (RD or TO) simulation cell in order to reduce periodicity effects, as has been suggested previously.²⁹ First of all, it is clear that the periodicity effects when taking the full simulation cell into account are as strong for all three cell geometries (Figure 5), albeit not identical for a given l . We note, however, that the influence from the periodicity on the

quadrupolar ($l = 2$) fluctuations is significantly lower in the RD and TO geometries than for the cube. Since the magnitude of the boundary effect for a given l decays as $R^{-(2l+1)}$, this means that the “leading-order” contribution to the boundary effects is about 50% smaller (Figure 6) in the RD and TO geometries compared to when using a cubic simulation cell. On the other hand, which simulation cell to be used also depends on the specific system under study. For example, if one wants to simulate a macromolecule (e.g., a protein) with a particularly large molecular octupole moment, a cubic box may be the most appropriate one, due to its lower lattice coupling for the system octupole moment. Furthermore, it may also be advantageous, when using rotational constraints, to orient the axis of the largest electrostatic moment along the axis with the lowest value of $S_{lm}^{(X)}$; for example, orienting the octupole moment in the $S_{33}^{(R)}$ or $S_{33}^{(I)}$ direction gives a lattice interaction of less than 3% of the value obtained when the octupole is oriented along the $S_{32}^{(I)}$ axis.

The present study provides a quantitative understanding of the isotropic and anisotropic parts of the solvation in a polar system under PBCs and puts them in relation to the behavior of an isotropic system. This understanding is essential for the possibility to remedy the periodicity effects, for example by imposing suitable bias functions in an MC simulation in order to remove the anisotropic self-interaction.

APPENDIX A. DERIVATION OF EQ 18

Inserting the definition from eq 16 into the left-hand-side of eq 18 and performing the straightforward summation over X yields

$$\sum_{m,X} S_{lm}^{(X)} = \sum_{m=0}^l (2 - \delta_{m0}) (-1)^{l+m} f(l, l, m, -m) \times \sum_{n \neq 0} \frac{C_{2l,0}^R(\Omega)}{|an|^{2l+1}} = \sum_{m=-l}^l (-1)^{l+m} f(l, l, m, -m) \sum_{n \neq 0} \frac{C_{2l,0}^R(\Omega)}{|an|^{2l+1}} \quad (29)$$

where we have used the fact that $f(l, l, m, -m) = f(l, l, -m, m)$. Inserting eq 17 and the expression for the $3j$ symbol⁴⁴ gives after some simplifications

$$\sum_{m,X} S_{lm}^{(X)} = \left(\sum_{n \neq 0} \frac{C_{2l,0}^R(\Omega)}{|an|^{2l+1}} \right) \times \sum_{m=-l}^l (-1)^{l+m} \times \frac{(2l)!}{(l+m)!(l-m)!} = \left(\sum_{n \neq 0} \frac{C_{2l,0}^R(\Omega)}{|an|^{2l+1}} \right) \times \sum_{m=-l}^l (-1)^{l+m} \binom{2l}{l+m} \quad (30)$$

where we have used the combinatorial definition of the binomial coefficients and the fact that the lattice sum is m -independent. We now make the change of summation index $l+m \rightarrow m'$, leading to

$$\sum_{m,X} S_{lm}^{(X)} = \left(\sum_{n \neq 0} \frac{C_{2l,0}^R(\Omega)}{|an|^{2l+1}} \right) \times \sum_{m'=0}^{2l} (-1)^{m'} \binom{2l}{m'} = 0 \quad (31)$$

where we have used another standard relation for the binomial

coefficients.⁴⁴ We finally note that eq 18 is valid regardless of the lattice type.

AUTHOR INFORMATION

Corresponding Author

*E-mail: joakim.stenhammar@fkem1.lu.se.

ACKNOWLEDGMENT

We would like to thank Björn Linse for help in deriving eq 18. This work was financed by the Swedish Research Council (VR) through the Linnaeus Center of Excellence Organizing Molecular Matter (Grant No. 239-2009-6794).

REFERENCES

- (1) Frenkel, D.; Smit, B. *Understanding Molecular Simulation*, 2nd ed.; Academic Press: San Diego, CA, 2002.
- (2) Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids*; Oxford University Press: New York, 1989.
- (3) Valleau, J. P. In *The Problem of Long-Range Forces in the Computer Simulation of Condensed Media*; Ceperley, D., Ed.; NRCC Proceedings 1980, Vol. 9, pp 3–8.
- (4) Hummer, G.; Pratt, L.; Garcia, A.; Neumann, M. In *Simulation and Theory of Electrostatic Interactions in Solution*; Pratt, L. R., Hummer, G., Eds.; AIP Conference Proceedings 1999, Vol. 492, pp 84–103.
- (5) Karttunen, M.; Rottler, J.; Vattulainen, I.; Sagui, C. *Curr. Top. Membr.* **2008**, *60*, 49–89.
- (6) Ewald, P. *Ann. Phys.* **1921**, *64*, 253–287.
- (7) Hockney, R. W.; Eastwood, J. W. *Computer Simulations Using Particles*; McGraw-Hill: New York, 1981.
- (8) Darden, T. A.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993**, *98*, 10089–10092.
- (9) Essmann, U.; Perera, L.; Berkovitz, M. L.; Darden, T. A.; Lee, H.; Pedersen, L. *J. Chem. Phys.* **1995**, *103*, 8577–8593.
- (10) Deserno, M.; Holm, C. *J. Chem. Phys.* **1998**, *109*, 7678–7693.
- (11) Valleau, J. P.; Whittington, S. G. In *Statistical Mechanics. Part A: Equilibrium Techniques*; Berne, B. J., Ed.; Plenum: New York, 1977; Chapter 4.
- (12) Pratt, L. R.; Haan, S. W. *J. Chem. Phys.* **1981**, *74*, 1864–1872.
- (13) Pratt, L. R.; Haan, S. W. *J. Chem. Phys.* **1981**, *74*, 1873–1876.
- (14) Hummer, G.; Pratt, L.; Garcia, A. *J. Phys. Chem. A* **1998**, *102*, 7885–7895.
- (15) Hünenberger, P. H. In *Simulation and Theory of Electrostatic Interactions in Solution*; Pratt, L. R., Hummer, G., Eds.; AIP Conference Proceedings 1999; Vol. 492; pp 17–83.
- (16) Bergdorf, M.; Peter, C.; Hünenberger, P. *J. Chem. Phys.* **2003**, *119*, 9129–9144.
- (17) Kastenholz, M. A.; Hünenberger, P. H. *J. Chem. Phys.* **2006**, *124*, 224501.
- (18) Hünenberger, P. H.; McCammon, J. A. *Biophys. Chem.* **1999**, *78*, 69–88.
- (19) Kastenholz, M. A.; Hünenberger, P. H. *J. Phys. Chem. B* **2004**, *108*, 774–788.
- (20) de Vries, A.; Chandrasekhar, I.; van Gunsteren, W.; Hünenberger, P. *J. Phys. Chem. B* **2005**, *109*, 11643–11652.
- (21) Reif, M. M.; Krätler, V.; Kastenholz, M. A.; Daura, X.; Hünenberger, P. H. *J. Phys. Chem. B* **2009**, *113*, 3112–3128.
- (22) Boresch, S.; Steinhauser, O. *Ber. Buns. Phys. Chem.* **1997**, *101*, 1019–1029.
- (23) de Leeuw, S. W.; Perram, J. W.; Smith, E. R. *Proc. R. Soc. London* **1980**, *A 373*, 27–56.
- (24) Fröhlich, H. *Theory of Dielectrics*, 2nd ed.; Oxford University Press: London, 1958.
- (25) Jackson, J. D. *Classical Electrodynamics*, 3rd ed.; Wiley: New York, 1999.

- (26) Nijboer, B. R. A.; de Wette, F. W. *Physica* **1958**, *24*, 422–431.
- (27) de Wette, F. W.; Schacher, G. E. *Phys. Rev.* **1965**, *137*, A78–A91.
- (28) Stenhammar, J.; Linse, P.; Karlström, G. *J. Chem. Phys.* **2009**, *131*, 164507.
- (29) Adams, D. J. In *The problem of long-range forces in the computer simulation of condensed media*; Ceperley, D., Ed.; NRCC Proceedings 1980, Vol. 9, p 13.
- (30) Dzwiniel, W.; Kitowski, J.; Mościński, J. *Mol. Simul.* **1991**, *7*, 171–179.
- (31) Adams, D. *CCPS Info. Quart.* **1983**, *10*, 30–36.
- (32) Smith, W. *CCPS Info. Quart.* **1983**, *10*, 37–42.
- (33) Bekker, H. J. *Comput. Chem.* **1997**, *18*, 1930–1942.
- (34) Kusalik, P. G.; Mandy, M. E.; Svishchev, I. M. *J. Chem. Phys.* **1994**, *100*, 7654–7664.
- (35) Wassenaar, T. A.; Mark, A. E. *J. Comput. Chem.* **2006**, *27*, 316–325.
- (36) Smith, W.; Fincham, D. *Mol. Simul.* **1993**, *10*, 67–71.
- (37) Stenhammar, J.; Linse, P.; Malmqvist, P.-Å.; Karlström, G. *J. Chem. Phys.* **2009**, *130*, 124521.
- (38) Stenhammar, J.; Linse, P.; Wennerström, H.; Karlström, G. *J. Phys. Chem. B* **2010**, *114*, 13372–13380.
- (39) Böttcher, C. J. F. *Theory of Electric Polarisation*, 1st ed.; Elsevier: Amsterdam, 1952.
- (40) Scaife, B. K. P. *Principles of Dielectrics*; Oxford University Press: New York, 1998.
- (41) Neumann, M. *Mol. Phys.* **1983**, *50*, 841–858.
- (42) de Leeuw, S. W.; Perram, J. W.; Smith, E. R. *Annu. Rev. Phys. Chem.* **1986**, *37*, 245–270.
- (43) McQuarrie, D. A. *Statistical Mechanics*; University Science Books: Sausalito, CA, 2000.
- (44) *Digital Library of Mathematical Functions*, 2011-08-29; National Institute of Standards and Technology. <http://dlmf.nist.gov> (accessed October 2011).
- (45) Linse, P. *Molsim*, version 5; Lund University: Lund, Sweden, 2011.
- (46) Stenhammar, J.; Linse, P.; Karlström, G. *J. Chem. Phys.* **2010**, *132*, 104507.

Statistical Convergence of Equilibrium Properties in Simulations of Molecular Solutes Embedded in Lipid Bilayers

Chris Neale,^{†,‡} W. F. Drew Bennett,[§] D. Peter Tieleman,[§] and Régis Pomès^{*,†,‡}

[†]Molecular Structure and Function, The Hospital for Sick Children, 555 University Avenue, Toronto, Ontario, Canada M5G 1X8

[‡]Department of Biochemistry, University of Toronto, Toronto, Ontario, Canada

[§]Department of Biological Sciences, Institute for Biocomplexity and Informatics, University of Calgary, Calgary, Alberta, Canada

ABSTRACT: In recent years, atomistic molecular simulations have become a method of choice for studying the interaction of small molecules, peptides, and proteins with biological membranes. Here, we critically examine the statistical convergence of equilibrium properties in molecular simulations of two amino acid side-chain analogs, leucine and arginine, in the presence of a hydrated phospholipid bilayer. To this end, the convergence of the standard binding free energy for the reversible insertion of the solutes in the bilayer is systematically assessed by evaluating dozens of separate sets of umbrella sampling calculations for a total simulation time exceeding 400 μ s. We identify rare and abrupt transitions in bilayer structure as a function of solute insertion depth. These transitions correspond to the slow reorganization of ionic interactions involving zwitterionic phospholipid headgroups when the solutes penetrate the lipid–water interface and when arginine is forced through the bilayer center. These rare events are shown to constitute hidden sampling barriers that limit the rate of convergence of equilibrium properties and result in systematic sampling errors. Our analysis demonstrates that the difficulty of attaining convergence for lipid bilayer-embedded solutes has, in general, been drastically underestimated. This information will assist future studies in improving accuracy by selecting a more appropriate reaction coordinate or by focusing computational resources on those regions of the reaction coordinate that exhibit slow convergence of equilibrium properties.

INTRODUCTION

Biological membranes enable life by maintaining different environments within and between cells, protecting cellular machinery from harsh and dilute external environments,¹ and compartmentalizing eukaryotic cells.² The defining component of a biological membrane is the lipid bilayer, a bilamellar sheet of oriented amphipathic lipids that collectively sandwich a hydrophobic interior between two hydrophilic surfaces, thus creating a barrier to the passage of materials.³ While biological membranes are significantly more complex than neat lipid bilayers,^{4,5} understanding the properties of neat lipid bilayers and their interactions with simple molecular solutes is a key step to understanding the properties of biological membranes.

The structure and bulk properties of lipid bilayers have been investigated with theoretical methods including molecular dynamics (MD) simulations.^{6,7} In recent years, atomistic MD simulations have been used to compute bilayer properties^{8,9} and equilibrium distributions of solutes across lipid bilayers.^{10–17} While simulations are useful tools for determining the equilibrium properties of complex systems, all simulation studies are susceptible to sampling errors,¹⁸ particularly when simulation time scales are less than or similar to the autocorrelation times of the degrees of freedom (DOFs) on which observed values depend.¹⁹ In this case, statistical measures such as the mean, μ , and standard deviation, σ , of values observed in simulations change significantly with increased sampling. This time dependence is a hallmark of insufficient sampling. Two types of sampling errors may arise from insufficient sampling: statistical and systematic errors. Statistical sampling errors primarily affect a value's precision, causing oscillations about an average value as

the simulation time is increased. Conversely, systematic sampling errors directly affect a value's accuracy. When systematic sampling errors arise from insufficient sampling, statistical values drift unidirectionally toward the correct values as the equilibration time is increased and the systematic sampling errors become smaller.

In some cases, a separation of time scales exists whereby one DOF relaxes much more slowly than all of the other DOFs. One may then conduct a series of simulations restrained to different positions along this slowly relaxing DOF or reaction coordinate, thus reducing the amount of simulation time required for statistical properties to converge. One technique that makes use of this procedure is known as umbrella sampling (US).^{20,21} However, the convergence and accuracy of average properties computed from US simulations can also suffer from systematic sampling errors introduced by slow relaxation in DOFs orthogonal to the selected reaction coordinate.²² The barriers that underlie long autocorrelation times in orthogonal DOFs are referred to as hidden barriers because, although they retard convergence, they are not apparent along the reaction coordinate. Conversely, an explicit barrier lies along the reaction coordinate. A schematic representation of a hidden barrier is presented in Figure 1. In principle, the system outlined in Figure 1 could be sampled much more efficiently if the DOF with the largest autocorrelation time was chosen as the reaction coordinate for US. Unfortunately, optimal reaction coordinates are rarely known *a priori*, and reaction coordinates are thus often selected simply because they are physically intuitive. To compute

Received: May 6, 2011

Published: October 05, 2011

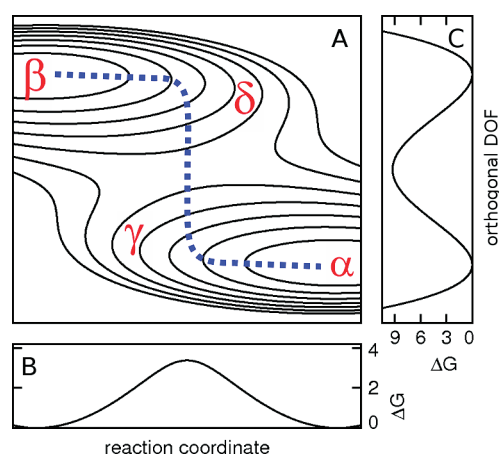


Figure 1. Schematic illustration of a free energy sampling barrier in a degree of freedom (DOF) orthogonal to the reaction coordinate. (A) Isoenergy contour representation of an example two-dimensional free energy surface. Two local minima designated α and β are connected by a dotted line that, as it progresses along the designated reaction coordinate (x axis), follows the path of lowest free energy in an orthogonal DOF (y axis). Two states that do not lie along this path are designated γ and δ . (B) Free energy profile along the reaction coordinate. (C) Free energy profile along the orthogonal DOF.

the equilibrium distribution of molecular solutes in lipid bilayers, one intuitive reaction coordinate is the distance between the solute and the bilayer center along the bilayer normal.

In this study, we examine the statistical convergence of the standard binding free energy for the immersion of two amino acid side-chain analogs in a lipid bilayer. First, we examine the partitioning of a chemically simple hydrophobic solute, methylpropane, the side chain analog of leucine. Second, we consider an amphipathic cationic solute, *n*-propylguanidinium, the side chain analog of arginine. These solutes are small and have few internal DOFs. This simplicity allows us to identify hidden barriers to solute insertion systematically. Moreover, these solutes are biologically relevant moieties whose solvation in a lipid bilayer may be relevant in the context of larger solutes, such as proteins.

For each solute, we use US to calculate the free energy profile or potential of mean force (PMF) governing the solute's axial probability distribution from bulk water to the center of a 1,2-dioleoyl-*sn*-glycero-3-phosphatidylcholine (DOPC) bilayer. These simulations extend two earlier studies of the distribution of hexane¹³ and the side chain analogs methylpropane and *n*-propylguanidinium.²³ We increase the respective sampling of methylpropane and *n*-propylguanidinium by factors of 80 and 40 and compute the PMF many times using different initial conditions. The increased simulation time and the systematic evaluation of the dependence of the binding free energy on initial conditions, both made possible by increases in available computer power, make it possible to quantify systematic sampling errors. Importantly, our in-depth analysis of statistical sampling convergence in simulations of lipid bilayers identifies the immersion depths at which these side chain analogs are particularly susceptible to systematic sampling errors. We elucidate the molecular underpinnings of these sampling errors and present a comprehensive view of the structural and thermodynamic bases of lipid solvation of these two biologically relevant solutes.

THEORY AND METHODS

Simulation Protocol. The simulation systems consisted of methylpropane or *n*-propylguanidinium in hydrated DOPC bilayers. MD simulations were conducted with version 4.0.7 of the GROMACS simulation package.²⁴ The water model was TIP3P.²⁵ Methylpropane and *n*-propylguanidinium were modeled by the OPLS-AA/L parameters^{26,27} for the side chains of leucine and arginine, respectively, where the α -carbon was replaced by a hydrogen atom and the charge on the β -carbon was adjusted to yield an integral molecular charge.²³ DOPC was modeled by the Berger parameters.²⁸ For combination with OPLS-AA/L solutes, the Coulombic 1–4 intramolecular interactions of DOPC were reduced to half magnitude in spite of the fact that self-consistent combination of the Berger and OPLS-AA/L parameter sets is now possible using the half- ϵ double-pairlist method.²⁹ Lennard-Jones interactions were evaluated using a group-based cutoff and truncated at 1 nm without a smoothing function. Coulomb interactions were calculated using the smooth particle-mesh Ewald method^{30,31} with a real-space cutoff of 1 nm and a Fourier grid spacing of 0.12 nm. Simulation in the *NpT* ensemble was achieved by semi-isotropic coupling to Berendsen barostats³² at 1 bar with coupling constants of 1 ps and coupling the water, lipids, and solute to three separate Berendsen thermostats³² at 298 K with coupling constants of 0.1 ps as in the previous study.²³ Bonds involving hydrogen atoms were constrained with SETTLE³³ and P-LINCS³⁴ for water and other molecules, respectively. The integration time step was 2 fs. The nonbonded pairlist was updated every 20 fs. Coordinates were saved every 10 ps.

System Setup and Umbrella Sampling. A configuration containing a DOPC lipid bilayer with 32 lipids per leaflet was obtained from the methylpropane-in-DOPC simulations of MacCallum et al.²³ To increase the spatial separation of the lipid bilayers across the periodic boundary, the simulation box was elongated along the bilayer normal in the Cartesian z dimension. Excess water was added to this newly-created cavity, resulting in a total of 4555 water molecules in the entire system. This DOPC bilayer was simulated for 108 ns in the absence of any solute, during which the area per lipid (APL; see Table 1 for acronyms and symbols used throughout this article) did not drift systematically (data not shown). Over these 108 ns, the APL was 0.644 ($\sigma = 0.01$) nm² and the spatial extent of the box along z was 10.8 ($\sigma = 0.2$) nm. (Throughout this article, standard deviations of the sample, $\sigma = [1/(N-1)\sum_{i=1}^N(v_i - \mu)^2]^{1/2}$, for N values of the sample, v , with mean μ are shown in parentheses while the \pm symbol is reserved for standard deviations of the mean, $\sigma_M = [1/(M-1)\sum_{j=1}^M(\mu_j - \bar{\mu})^2]^{1/2}$, for M estimates of the mean, μ , with overall mean $\bar{\mu}$. We have chosen to display the standard deviation of the mean instead of the standard error of the mean because the former reflects the range of values that can be expected for the mean from a single set of US simulations, which is currently the most common application of US.) Coordinates were extracted after 8, 48, and 108 ns for use in US simulations. A solute was embedded in each of these bilayer conformations using the inflategro routine.³⁵ This solute insertion process was repeated 65 times while varying the axial position of the center of mass (COM) of the solute relative to that of the lipid bilayer from $z = -3.2$ nm to $+3.2$ nm in 0.1 nm increments, where z represents the solute immersion depth in the bilayer. This entire protocol was then repeated, for each of the three bilayer conformations and each of 65 axial coordinates, after translating the bilayer by

Table 1. Selected Acronyms and Symbols Used Throughout This Article

symbol	definition
APL	area per lipid
COM	center of mass
SDF	spatial distribution function
σ_M	standard deviation of the mean
t_{eq}	simulation time discarded as equilibration
z	the Cartesian z axis
z	distance along z from the bilayer COM to the solute COM (solute immersion depth)
z_i^0	center of the harmonic restraining potential along z for umbrella i
ΔG_z	free energy as a function of z
$\Delta G_{\text{bind}}^\circ$	standard binding free energy
$\rho_{\text{H}_2\text{O}}$	water density profile across a neat bilayer
ρ_{P}	lipid phosphorus density profile across a neat bilayer
N_{wat}	number of water oxygen atoms in the solute's first solvation shell
N_{P}	number of lipid phosphorus atoms in the solute's first solvation shell
r_{P}	distance between lipid phosphorus atom and the solute COM in the xy plane
$z_{\text{P}}^{\text{upper}}$	lipid phosphorus atom height deviation in upper leaflet (not necessarily the proximal leaflet) from their mean position in a neat bilayer
$\Delta_{\text{acyl}}^{\text{upper}}$	deviation in acyl chain order parameters in the upper leaflet (not necessarily the proximal leaflet) from their mean values in a neat bilayer
θ	angle (used to report the solute orientation)

half a box length in the bilayer plane. This translation doubled the number of uncorrelated bilayer surfaces used to determine the PMF of solute insertion. In the methylpropane simulations, the solute was embedded in two different initial orientations, with the solute $\text{C}_\gamma\text{H}_\gamma$ vector aligned to the z axis or perpendicular to it. Note that solute atomic nomenclature is based on the cognate amino acid side chain. In the *n*-propylguanidinium simulations, the all-*trans* solute was embedded in three different initial orientations, with the $\text{C}_\delta\text{C}_\zeta$ vector of the guanidino group pointing either up or down along the bilayer normal or along the bilayer plane. Considering that each bilayer has two leaflets, the methylpropane and *n*-propylguanidinium simulations yield, respectively, a total of 24 and 36 distinct sets of US simulations for separate evaluations of the PMF from bulk water to the center of the bilayer.

The primary objective of these simulations is to evaluate the rate at which free energies attain convergence in US simulations of solute insertion along the lipid bilayer normal. Thus, we have not evaluated the effect of salt concentration on the mechanism of solute insertion. In addition, we have deliberately avoided a potential source of quasi-nonergodic sampling, namely, the slow equilibration of the distribution of a single counterion such as Cl^- . To this end, we did not add a counterion to the *n*-propylguanidinium systems, which retain a +1 net charge. Because the *n*-propylguanidinium PMF is determined not only by strong Coulombic interactions between charged lipid headgroups and the solute but also, indirectly, by the network of ionic interactions between the headgroups themselves, it is unlikely that the addition of a single counterion would have a significant effect on the outcome of this study (see the Results). Although there has been much debate about the protonation state of arginine within a lipid bilayer^{36,37} and it is possible to compute $\text{p}K_a$ profiles for charged residues from molecular simulations,^{38–40} we assume that *n*-propylguanidinium remains protonated, and we focus on the statistical sampling convergence of the bilayer interacting with *n*-propylguanidinium in its cationic state.

US simulations were conducted for 205 ns for each initial conformation and under the influence of each restraining potential (umbrella). During these simulations, the solute insertion

depth, z , was harmonically restrained to a specified value, z_i^0 , in each umbrella i , with a force constant of 3000 kJ/mol/nm^2 , and was stored every 1 ps. While we did not systematically evaluate the effect of varying the spacing or the force constant of our umbrella potentials on the PMF, we did ensure that there was sufficient overlap between adjacent histograms along z to permit the computation of PMFs (data not shown).

Finally, the methylpropane and *n*-propylguanidinium US simulations of MacCallum et al.²³ were extended to 205 ns per umbrella using pre-existing restart files and GROMACS 4.0.7.

Methodological Comparison to Previous Simulations. The simulations reported in this study were conducted under similar conditions to those of MacCallum et al.,²³ with seven differences: (i) We used GROMACS version 4.0.7²⁴ in place of version 3.3.1.⁴¹ We also used (ii) a greater number of water molecules, (iii) different starting conformations, and (iv) a different method to embed solute molecules at specified bilayer depths. (v) Our systems contained a single solute molecule rather than the two molecules separated by 3.7 nm along the bilayer normal as employed earlier.²³ (vi) Our *n*-propylguanidinium simulations did not contain a counterion. Finally, (vii) our simulations were considerably longer. We simulated methylpropane for a total of 160 μs and *n*-propylguanidinium for a total of 240 μs compared to earlier totals of 2 and 6 μs , respectively.²³ The water model was TIP3P in both studies in spite of the fact that MacCallum et al. indicated using the SPC water model⁴² in their publication.²³

Free Energies and Standard States. The values of z sampled by each solute in the US simulations were converted to PMFs using Alan Grossfield's implementation⁴³ of the weighted histogram analysis method (WHAM).⁴⁴ To this end, recorded values of the solute insertion depth in the range $-3.25 \leq z \leq 3.25 \text{ nm}$ were distributed among 2600 histogram bins, and the WHAM calculation was performed with a tolerance of 1×10^{-5} . This was done separately for each set of US simulations. Each resulting PMF describes the free energy as a function of solute immersion depth, ΔG_z , from bulk water ($z = -3.25 \text{ nm}$) across the bilayer center ($z = 0 \text{ nm}$) to bulk water ($z = 3.25 \text{ nm}$). Exploiting the

symmetry of the system with respect to the $z = 0$ nm plane, we present the PMF for the absolute value of z . Each PMF was then shifted such that the average value of ΔG_z in the range $3 < z \leq 3.25$ nm equaled zero. Finally, the standard binding free energy, $\Delta G_{\text{bind}}^\circ$, was determined by trapezoid integration of this PMF according to

$$e^{-\beta\Delta G_{\text{bind}}^\circ} = \frac{\int_{z_{\text{min, bound}}}^{z_{\text{max, bound}}} e^{-\beta\Delta G_z} dz}{\int_{z_{\text{min, unbound}}}^{z_{\text{max, unbound}}} e^{-\beta\Delta G_z} dz} \quad (1)$$

where $\beta = (k_B T)^{-1}$, k_B is the Boltzmann constant, and T is the absolute temperature. Additionally, $z_{\text{min, bound}}$ and $z_{\text{max, bound}}$ define the bound state and, likewise, $z_{\text{min, unbound}}$ and $z_{\text{max, unbound}}$ define the unbound state. We assign a constant value of 0 kcal/mol to the PMF in the unbound state. Equation 1 is integrated over the range Δz :

$$\begin{aligned} \Delta z &= z_{\text{max, bound}} - z_{\text{min, bound}} \\ &= z_{\text{max, unbound}} - z_{\text{min, unbound}} \end{aligned} \quad (2)$$

which ensures that the available volume is equal in the bound and unbound states,⁴⁵ thus calculating $\Delta G_{\text{bind}}^\circ$ on the basis of a volume-fraction partition coefficient.⁴⁶ Integrating over a single leaflet of the bilayer, we set $z_{\text{min, bound}} = 0$ nm and $z_{\text{max, bound}} = 3$ nm, which is beyond the point at which the solute becomes fully hydrated and the mean force acting on it vanishes (see the Results). Equation 2 ensures that eq 1 provides a standard-state evaluation of the binding free energy, since the relative bound and unbound probabilities depend on the volume of space available to the solute in these respective states. Thus, doubling the thickness of the water layer in the integral appearing in the denominator of eq 1 would double the population of the unbound state. Note that since we are integrating over z , this volume dependence reduces to a length dependence.

As shown above, the PMF allows definition of the bound state in a rational manner, by placing the boundary at the point where the mean force becomes zero. Nevertheless, sampling errors can lead to different definitions of the bound state in repeat simulations. While an increase in Δz has the same additive effect on the probabilities of the bound and unbound states, the already large probability of the bound state (see the Results) dictates that enlargement of the bound and unbound state definitions affects the unbound state integral (denominator in eq 1) much more dramatically than the bound state integral (numerator in eq 1). Considering the range of possible definitions of $z_{\text{max, bound}}$ based on the point where the mean force becomes zero (3.0 nm in this work and 3.7 nm in MacCallum et al.²³), $\Delta G_{\text{bind}}^\circ$ has an additional uncertainty component of $\beta^{-1} \ln(3.7/3.0) = 0.12$ kcal/mol. Confining the bound state to a hydrophobic length of 2.0 nm per leaflet, as some authors have done,⁴⁷ increases this uncertainty to $\beta^{-1} \ln(3.7/2.0) = 0.36$ kcal/mol, and even larger uncertainties are possible with larger solutes for which the mean force becomes zero at very large values of $z_{\text{max, bound}}$.

Unrestrained Simulations. To assess the binding mechanism directly, we conducted 10 unrestrained simulations of methylpropane in the presence of a DOPC bilayer. To this end, initial coordinates of the hydrated bilayer were extracted from the 108-ns simulation of a neat DOPC bilayer, and 10 methylpropane molecules were placed in bulk water. After energy minimization, 10 205-ns MD trajectories with different initial

velocities were generated in the absence of restraints. Note that these simulations were not used to evaluate the binding free energy.

Additional US Simulations at the Lipid–Water Interface.

To probe for the presence of a hidden sampling barrier at the lipid–water interface, we conducted 20 additional 180-ns US simulations of methylpropane with an umbrella at $z_i^0 = 2.0$ nm. Ten distinct starting conformations were drawn from unrestrained simulations in which a methylpropane molecule located near $z = 2.0$ nm subsequently progressed deeply into the hydrophobic interior. An additional 10 starting conformations in which methylpropane molecules located near $z = 2.0$ nm did not subsequently bind the bilayer but instead diffused back into bulk water were also selected.

Data Analysis. The number of water molecules in the first solvation shell of the solutes, $N_{\text{H}_2\text{O}}$, was calculated on the basis of a heavy-atom cutoff distance of 0.435 nm. This distance corresponds to the first minimum in the radial distribution function (RDF) of water oxygen atoms around methylpropane heavy atoms in aqueous solution (data not shown). The cutoff for determining the number of phosphorus atoms interacting closely with the solutes, N_p , was 0.585 nm. Because the corresponding RDF is complex (data not shown), we chose to simply increase the aqueous cutoff by 0.15 nm, approximately the length of the P–O bond in the lipid phosphate group. Axial density profiles for water oxygen atoms, $\rho_{\text{H}_2\text{O}}$, and for lipid headgroup phosphorus atoms, ρ_p , were calculated from the 108-ns simulation of a neat DOPC bilayer with the GROMACS `g_density` tool (correcting for fluctuations in the length of the box along z by modifying the `g_density` algorithm to bin the data outward from the bilayer center) after centering the bilayer along z using the GROMACS `trjconv` tool (note that the centering algorithm was modified to center the COM rather than the mean of the maximum and minimum values). For ease of comparison, density profiles were scaled such that their maximum value is numerically similar to the maximum $N_{\text{H}_2\text{O}}$ for that solute. The distances in the xy plane between the solute COM and each headgroup phosphorus atom, r_p , were measured with the GROMACS `g_dist` tool. The distances along z between the bilayer COM and the headgroup phosphorus atom of lipids in the upper leaflet were measured with the GROMACS `g_dist` tool, and deviations of the measured values from the mean distance in a neat DOPC bilayer (2.01 nm) are represented by z_p^{upper} , which, for some figures, was computed only for lipids with $r_p \leq 1$ nm. Order parameters were computed for all saturated nonterminal acyl-chain carbon atoms by reconstructing hydrogen atom positions assuming tetrahedral geometry using the GROMACS `g_order` tool, as outlined previously,⁷ for those lipids with $r_p \leq 1$ nm. These order parameters were averaged for each chain position in each simulation, and the deviations of these mean values from reference values obtained from simulations of a neat bilayer were computed for each chain position and then grouped together as $\Delta S_{\text{acyl}}^{\text{upper}}$.

As we did for PMFs, most data are shown as a function of the absolute value of z in order to enhance statistics. The only exception is the solute orientation, which was evaluated as the angle, θ , between the positive bilayer normal and the $C_\gamma H_\gamma$ vector of methylpropane or the $C_\delta C_\zeta$ vector of *n*-propylguanidinium, and was processed as a function of the signed value of z to provide a measure of convergence.

Spatial distribution functions (SDFs) were created with the GROMACS `g_spatial` tool with a bin width of 0.05 nm after centering the bilayer COM at $z = 0$ nm and the solute COM

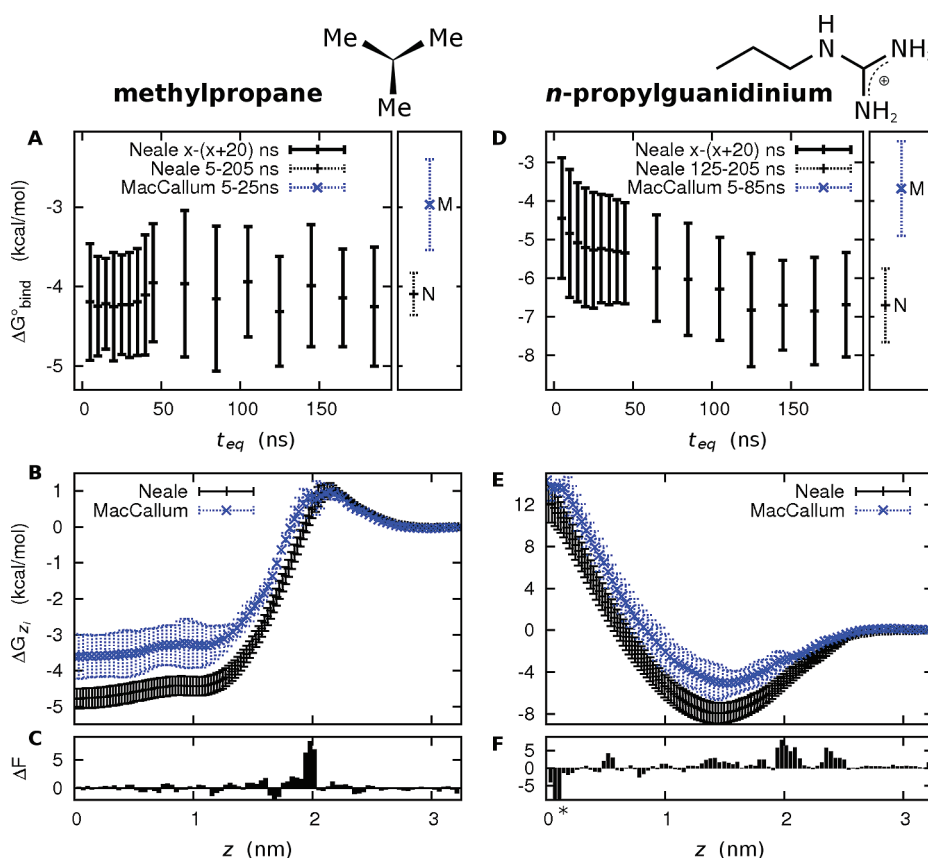


Figure 2. Free energies of (left) methylpropane and (right) *n*-propylguanidinium interacting with a DOPC bilayer. Chemical structures of each solute are displayed beside the titles. (A and D) $\Delta G_{\text{bind}}^{\circ}$ and the standard deviation of the mean, σ_M , based on a 20 ns sample per umbrella after discarding an increasing amount of simulation time, t_{eq} , as equilibration. The two values with dotted error bars represent $\Delta G_{\text{bind}}^{\circ}$ from full-production sampling in (black “+” labeled “N”) this work and (blue “x” labeled “M”) MacCallum et al.²³ (refer to Theory and Methods for the definition of full-production sampling). (B and E) The PMF and its σ_M for the solute from bulk water to the bilayer center from full-production sampling in (black “+” with solid error bars) this work and (blue “x” with broken error bars) MacCallum et al.²³ (C and F) The slope of the PMF from MacCallum et al.²³ is subtracted from the slope of the PMF calculated in this work to yield ΔF in kcal/mol/nm. (*) In part F, ΔF values extend to -20 kcal/mol/nm at $z = 0.1$ nm.

at $x = y = 0$ nm and then, to enhance resolution, repetitively concatenating the simulation trajectory data rotated about z from 10° through 360° in 10° increments. Molecular visualizations were prepared with VMD.⁴⁸ In figure captions, the term “full-production sampling” denotes the following sample per umbrella: for methylpropane, 5–205 ns in this work and 5–25 ns in MacCallum et al.,²³ and, for *n*-propylguanidinium, 125–205 ns in this work and 5–85 ns in MacCallum et al.²³ All error bars in figures represent σ_M , the standard deviation of the mean.

RESULTS

Solute Distribution Across a DOPC Bilayer. In this section, we analyze the statistical convergence of the standard binding free energy, $\Delta G_{\text{bind}}^{\circ}$, for the insertion of methylpropane and *n*-propylguanidinium in a DOPC bilayer based on PMF profiles along the bilayer normal. The PMF profiles were computed from US simulations with 205 ns at each of 33 umbrellas from bulk water to the bilayer center, repeated 24 and 36 times for methylpropane and *n*-propylguanidinium, respectively. These results are compared to those of previous US simulations with 25 and 85 ns per umbrella for methylpropane and *n*-propylguanidinium, respectively, each performed twice.²³ $\Delta G_{\text{bind}}^{\circ}$ was computed by integrating the PMF in the standard state (see Theory and Methods). To

select the sampling time range on which the PMFs were based, we used block averaging⁴⁹ to identify the duration of visible initial systematic sampling errors from our simulations and excluded that data from the computation of our PMFs. The PMFs from previous studies were taken from ref 23 without any modification.

Methylpropane. The value of $\Delta G_{\text{bind}}^{\circ}$ for methylpropane binding to a DOPC bilayer is shown as a function of equilibration time, t_{eq} , in Figure 2A. The average $\Delta G_{\text{bind}}^{\circ}$ computed on the basis of 20 ns per umbrella remains constant at -4.12 ± 0.14 kcal/mol with increasing t_{eq} between 5 and 185 ns (Table 2 and Figure 2A). Accordingly, the value of $\Delta G_{\text{bind}}^{\circ}$ computed on the basis of 5–205 ns per umbrella, -4.10 ± 0.27 kcal/mol, is within this range (Table 2 and Figure 2A). In comparison, the average value of $\Delta G_{\text{bind}}^{\circ}$ determined from the two previous simulations with less sampling and different starting conformations,²³ -2.97 ± 0.57 kcal/mol, differs from the average value obtained in this study by 1.1 kcal/mol (Table 2 and Figure 2A), and both absolute values of $\Delta G_{\text{bind}}^{\circ}$ from the two previous simulations are smaller than the smallest of 24 independent evaluations of $\Delta G_{\text{bind}}^{\circ}$ obtained in this work (Table 2). Importantly, the extension of the previously published simulations yields $\Delta G_{\text{bind}}^{\circ} = -3.96 \pm 0.29$ kcal/mol when computed on the basis of 5–205 ns per umbrella, comparable with the average value of $\Delta G_{\text{bind}}^{\circ}$ obtained from 24 PMFs of similar length in this work (Table 2).

Table 2. Standard Binding Free Energies of Methylpropane and *n*-Propylguanidinium to a DOPC Bilayer

sample (ns)	$\Delta G_{\text{bind}}^{\circ}$	component values
	(kcal/mol)	
	$\mu \pm \sigma_M$	
(i) methylpropane		
$\{x \rightarrow (x + 20)\} x = \{5, 25, 45, \dots, 185\}$	-4.12 ± 0.14^a	not shown
5 \rightarrow 205	-4.10 ± 0.27	-4.80, -4.41, -4.39, -4.37, -4.32, -4.29, -4.28, -4.22, -4.22, -4.21, -4.16, -4.16, -4.07, -4.02, -4.01, -3.99, -3.87, -3.87, -3.86, -3.83, -3.82, -3.75, -3.71, -3.70
5 \rightarrow 25 ^c	-2.97 ± 0.57	-3.37, -2.57
5 \rightarrow 205 ^d	-3.96 ± 0.29	-4.16, -3.76
(ii) <i>n</i> -propylguanidinium		
$\{x \rightarrow (x + 20)\} x = \{125, 145, 165, 185\}$	-6.77 ± 0.08^b	not shown
125 \rightarrow 205	-6.71 ± 0.96	-9.12, -9.03, -8.87, -8.56, -8.25, -8.11, -8.00, -7.98, -7.92, -7.51, -7.49, -7.47, -7.39, -7.10, -7.07, -7.04, -6.90, -6.75, -6.57, -6.47, -6.47, -6.31, -6.29, -5.96, -5.86, -5.85, -5.80, -5.66, -5.49, -5.43, -5.39, -5.33, -4.89, -4.39, -4.32, -4.00
5 \rightarrow 85 ^c	-3.68 ± 1.23	-4.56, -2.81
125 \rightarrow 205 ^d	-4.69 ± 0.01	-4.70, -4.69

^a Here, σ_M is computed from 10 blocks of 20 ns, each of which is first averaged over 24 PMFs. ^b Here, σ_M is computed from 4 blocks of 20 ns, each of which is first averaged over 36 PMFs. ^c MacCallum et al.²³ (2 PMFs). ^d Simulations extended from MacCallum et al.²³ (2 PMFs).

The largest difference between the PMFs for methylpropane insertion into a DOPC bilayer derived in the present and the previous study²³ occurs for solute insertion depths $1.85 \leq z \leq 2.05$ nm (Figure 2B,C), which correspond to the lipid–water interface (see below). At these depths, the PMF computed in this study smoothly decreases toward the global minimum located at the bilayer center, whereas the previous PMF does not.

n-Propylguanidinium. The value of $\Delta G_{\text{bind}}^{\circ}$ for *n*-propylguanidinium binding to a DOPC bilayer is shown as a function of t_{eq} in Figure 2D. The value of $\Delta G_{\text{bind}}^{\circ}$ drifts systematically as t_{eq} increases between 5 and 125 ns per umbrella. It is only after discarding the initial 125 ns of simulation that the value of $\Delta G_{\text{bind}}^{\circ}$ computed on the basis of 20 ns per umbrella converges to -6.77 ± 0.08 kcal/mol (Table 2 and Figure 2D). This estimate is consistent with the value of -6.71 ± 0.96 kcal/mol based on 125–205 ns per umbrella (Table 2 and Figure 2D). The estimate of $\Delta G_{\text{bind}}^{\circ}$ determined from the previous simulations,²³ -3.68 ± 1.23 kcal/mol, differs from the average value obtained in this study by 3 kcal/mol (Table 2 and Figure 2D). The extension of the previous simulations yields an estimate of $\Delta G_{\text{bind}}^{\circ} = -4.69 \pm 0.01$ kcal/mol based on 125–205 ns per umbrella. This value still differs by 2 kcal/mol from the average value of $\Delta G_{\text{bind}}^{\circ}$ obtained from the 36 PMFs based on new simulations of similar length (Table 2).

The global minimum of the PMF is located at insertion depths $1.4 \leq z \leq 1.5$ nm (Figure 2E). The largest differences with the previous simulations occur near insertion depths $z = 2.4, 2.0,$ and 0.1 nm (Figure 2E,F). At these depths, the slope of the new PMF profile is steeper than previously computed.²³

To ensure that our results were not influenced by pressure artifacts resulting from simulating a periodic system with net charge,⁵⁰ we evaluated the APL from simulations in which either methylpropane or *n*-propylguanidinium was restrained in bulk water at $z_i^0 = \pm 3.2$ nm based on a time-range of 125–205 ns per umbrella. The APL values were 0.642 ± 0.006 ($\sigma = 0.014$) nm² and 0.641 ± 0.005 ($\sigma = 0.014$) nm² for methylpropane and for *n*-propylguanidinium, respectively. It appears, therefore, that the APL in these simulations was not affected by the presence of a net charge.

Bilayer Reorganization. In this section, we analyze the structural properties underlying the slow convergence of

$\Delta G_{\text{bind}}^{\circ}$ identified in Figure 2. Solute insertion induces two types of structural perturbation of the lipid bilayer, which involve either (i) depression or (ii) protrusion of the bilayer surface. Depression occurs when the phosphate groups of lipids near the solute retract toward the bilayer center (Figure 3B,C,G,H) as acyl chains become more disordered (Figure 3D,I). Depression of the bilayer surface leaves the solute hydrated to a greater extent than expected on the basis of the water density profile across a neat bilayer, $\rho_{\text{H}_2\text{O}}$ (Figure 3A,F). Conversely, a protrusion of the bilayer surface occurs when the phosphate groups of lipids near the solute extend away from the bilayer center (Figure 3B,C,G,H) as acyl chains become more ordered (Figure 3D,I) and, as a result, the solute is hydrated to a lesser extent than expected on the basis of $\rho_{\text{H}_2\text{O}}$ (Figure 3A,F). These patterns of bilayer reorganization are apparent in the time-averaged spatial distribution functions (SDFs) depicted in Figure 4 and the simulation snapshots displayed in Figure 5.

To describe bilayer reorganization as a function of z , we conceptually divide a neat DOPC bilayer into four regions: (i) bulk water for $z > 2.5$ nm, (ii) the aqueous interface between water and phosphate groups spanning $2.0 < z \leq 2.5$ nm, (iii) the dry transition region between phosphate groups and acyl chains spanning $1.5 < z \leq 2.0$ nm, and finally, (iv) the hydrophobic region of the bilayer for $z \leq 1.5$ nm. Although this model differs from previous decompositions⁵¹ of the axial distribution of hydrated bilayers, it is pertinent to the present analysis.

Methylpropane. When methylpropane resides at the aqueous interface, the bilayer surface excludes the solute by forming a depression (Figure 3A–D). As a corollary, the number of lipid headgroup phosphorus atoms that closely interact with the solute is maximal in this region (Figure 3A). This type of lipid reorganization is depicted by the SDF in Figure 4A.

By contrast, when methylpropane is located in the dry transition region, the bilayer surface protrudes to include the solute (Figures 3A–D), as depicted by the SDF in Figure 4B. Thus, the local depression of the water–headgroup interface induced by the presence of the hydrophobic solute at the aqueous interface (Figures 4A, 5A) becomes a protrusion (Figures 4B, 5B) once the solute pierces the surface formed

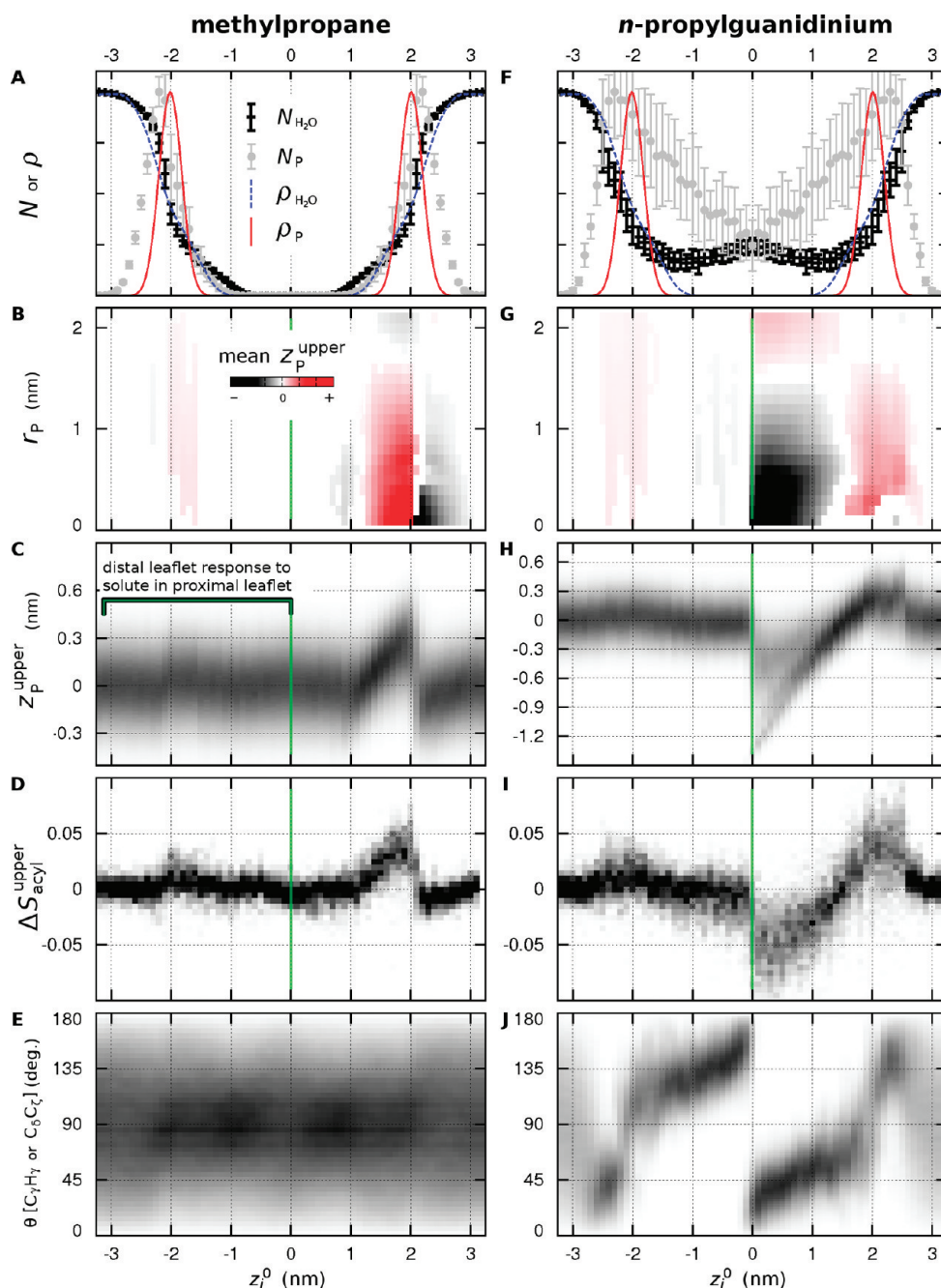


Figure 3. Structural reorganization of the bilayer as a function of restraint position, z_i^0 , for (left) methylpropane and (right) *n*-propylguanidinium partitioning into a DOPC bilayer via US. (A and F) Solvation of the solute by (black error bars) water, $N_{\text{H}_2\text{O}}$, and (gray error bars) lipid phosphorus atoms, N_{P} . Overlaid on this plot are the density profiles in a neat DOPC bilayer for (broken blue line) water, $\rho_{\text{H}_2\text{O}}$, and (solid red line) lipid phosphorus atoms, ρ_{P} . All profiles are scaled to the same maximum value. (B and G) Mean values of lipid height deviation in the upper leaflet, $z_{\text{P}}^{\text{upper}}$, are plotted for (red) positive and (black) negative displacements against (x axis) z_i^0 and (y axis) proximity of the lipid to the solute in the bilayer plane, r_{P} . The values of $z_{\text{P}}^{\text{upper}}$ in B and G were filtered by setting the mean value of $z_{\text{P}}^{\text{upper}} = 0$ whenever the standard deviation of the mean value of $z_{\text{P}}^{\text{upper}}$ included zero. (C and H) Probability distributions of $z_{\text{P}}^{\text{upper}}$ as a function of z_i^0 for lipids with $r_{\text{P}} < 1$ nm. (D and I) Distributions of mean values of the deviation in acyl chain order parameters in the upper leaflet from their mean values in a neat bilayer, $\Delta S_{\text{acyl}}^{\text{upper}}$, as a function of z_i^0 for lipids with $r_{\text{P}} < 1$ nm. (E and J) Probability distributions of solute orientation, θ , as a function of z_i^0 . In parts B–D and G–I, a vertical green line at $z_i^0 = 0$ nm indicates that the z axes of parts B and G and the y axes of parts C, D, H, and I represent changes that occur in a single leaflet as the solute traverses that leaflet (positive x axis values) and the opposing leaflet (negative x axis values). Histograms in parts C–E and H–J were normalized within each z_i^0 value.

by zwitterionic interactions involving phosphate (and choline) groups (Figure 3B,C).

Finally, as methylpropane enters the hydrophobic region of the bilayer, near $z = 1$ nm, it introduces a packing defect that

draws water molecules to a region that is normally dehydrated in a neat DOPC bilayer (Figure 3A), resulting in a slight retraction of nearby lipids toward the bilayer center (Figure 3B,C). Methylpropane is entirely dehydrated only for $z < 0.7$ nm (Figure 3A),

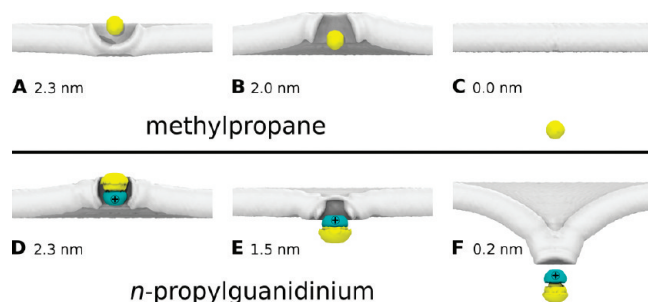


Figure 4. Spatial distribution functions (SDFs) of the solute and the proximal bilayer leaflet viewed from the side, looking along the plane of the bilayer. The SDFs depict (gray) phosphorus atoms in the proximal leaflet and (yellow) hydrophobic and (cyan) hydrophilic moieties of the solute for (A–C) methylpropane and (D–F) *n*-propylguanidinium. The z_i^0 value is reported alongside each SDF. Images are aligned vertically according to the bilayer COM. To facilitate solute visualization, the SDF of the phosphorus atoms depicts only the density that lies behind the COM of the solute.

where the solute does not affect the localization of the headgroups above it (Figures 3B, 3C, 4C, 5C).

Throughout the bilayer, methylpropane adopts nearly isotropic orientations whereby the angle between the positive bilayer normal and the $C_\gamma H_\gamma$ vector, θ , is sampled with a probability proportional to $\sin(\theta)$ (Figure 3E). There is a slight population shift toward orientations in which the $C_\gamma H_\gamma$ vector is parallel to the global bilayer plane when methylpropane is embedded in the bilayer, except at the bilayer center.

n-Propylguanidinium. When *n*-propylguanidinium resides at the aqueous interface, the bilayer surface protrudes to interact with the solute (Figure 3F–I). This reorganization is depicted by the time-averaged SDF in Figure 4D and illustrated by the snapshot in Figure 5D. In the aqueous limit of this region, at $z = 2.5$ nm, *n*-propylguanidinium is strongly oriented with the guanidino group pointing toward the bilayer center (Figure 3J). As the solute reaches $z = 2.0$ nm, a depth that corresponds to the peak in axial phosphate density in a neat DOPC bilayer (Figure 3F), the orientation of the solute is reversed, with the alkyl chain inserted in the nonpolar core and the guanidino group pointing away from the bilayer center (Figure 3J).

As *n*-propylguanidinium moves from the dry transition region to an immersion depth that corresponds to the hydrophobic core of a neat DOPC bilayer, the bilayer forms a depression that facilitates the continued interaction of the bilayer surface with the solute (Figure 3F–I). This reorganization is evident in the snapshot shown in Figure 5E and in the SDF depicted in Figure 4E, which correspond to the global minimum of the PMF, at $1.4 \leq z \leq 1.5$ nm (Figure 2E). The magnitude of the structural perturbation of the bilayer grows with deeper solute immersion (Figure 3F–I). Throughout the hydrophobic region of the bilayer, *n*-propylguanidinium remains partially hydrated and solvated by lipid headgroups (Figure 3F) and is strongly oriented with the guanidino group directed outward (Figure 3J). The extreme reorganization of the bilayer structure that occurs as the solute nears the bilayer center is evident in Figure 4F.

When *n*-propylguanidinium is restrained near the center of the bilayer, at $z = 0$ or 0.1 nm, bulk water may penetrate either leaflet (data not shown) and both polarized orientations are populated (Figure 3J).

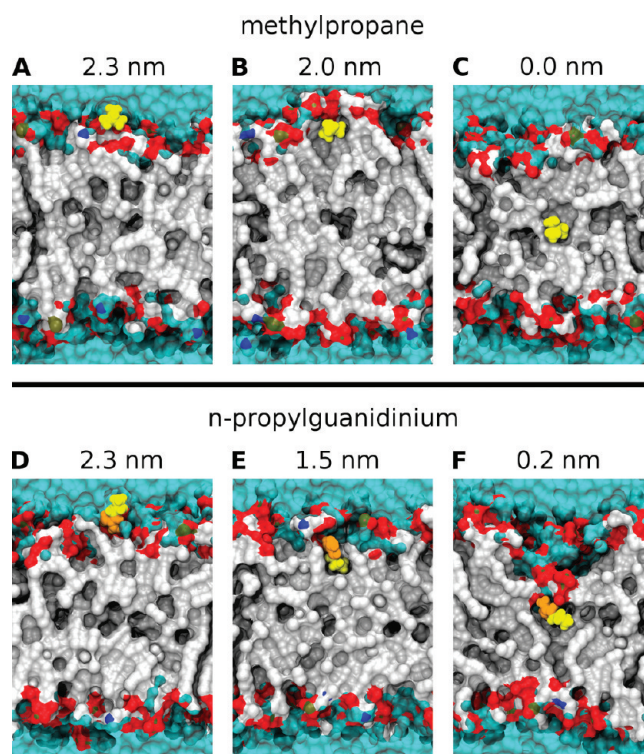


Figure 5. Snapshots of (A–C) methylpropane and (D–F) *n*-propylguanidinium after 200 ns of simulation. The z_i^0 value is reported alongside each snapshot. Color is added for (cyan) water, (white) lipid carbon, (red) lipid oxygen, (brown) lipid phosphorus, (blue) lipid nitrogen, (yellow) solute hydrophobic moiety, and (orange) solute hydrophilic moiety.

Structural Perturbations of the Opposing Leaflet. Significant changes in bilayer organization take place in both leaflets as the molecular solutes are embedded. We refer to the leaflet in which the solute is embedded as the proximal leaflet and to the opposing leaflet as the distal leaflet. As described above, the reorganization of the proximal leaflet predominantly occurs in the vicinity of the solute. In contrast, the reorganization of the distal leaflet is more global, with uniform displacements of the water–bilayer interface (Figure 3B,G) and uniform changes in order parameters (data not shown). As a result, the bilayer does not bend. Note that global bending modes of the bilayer may be inhibited to some extent by the presence of periodic boundary conditions in the x and y dimensions. Instead, local depressions and protrusions of the proximal interface induce respective increases in disorder and order of lipids throughout the distal leaflet (Figure 3D,I). These compensating changes in lipid ordering result in thinning and thickening of the distal leaflet, respectively, both of which uniformly displace the distal interface (Figure 3B,G). Compensating effects also extend to more distant regions of the proximal leaflet (Figure 3B,G) whenever the solute is restrained to values of z at which the PMF has a large slope (see Figure 2B,E). Throughout these processes, the change in average area per lipid remains smaller than 0.013 nm^2 (data not shown).

Effect of Restraints on the Sampling of Bilayer Perturbations. To assess the relevance of the US ensembles depicted in Figure 4 to the mechanism governing the partitioning of methylpropane into the bilayer, we conducted nonequilibrium simulations in the absence of restraints. Ten simulations, each

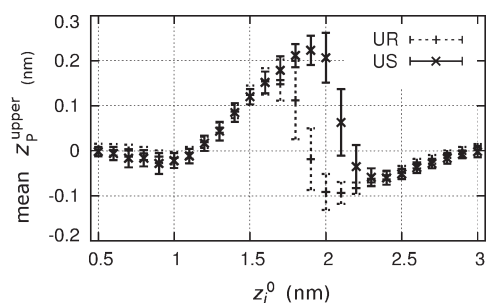


Figure 6. Bilayer reorganization as a function of methylpropane immersion sampled by different methods. The mean value of z_p^{upper} for lipids with $r_p < 1$ nm is shown from (“+” and broken lines denoted UR) unrestrained sampling and (“x” and solid lines) US. Error bars represent the standard deviation of the mean. The histogram bin width for unrestrained sampling is 0.1 nm.

containing 10 methylpropane molecules initially located in the aqueous phase, were run for 205 ns each. In the course of these simulations, an average of 6.4 ± 1.3 solute molecules spontaneously partitioned into the lipid bilayer, no unbinding events took place, and the distribution of methylpropane along the bilayer normal was still drifting with increasing simulation time when the simulations were stopped (data not shown). Binding events were characterized by long waiting times (90 ± 56 ns considering only the 64 binding events that occurred) followed by very rapid progression through the transition state to the bound state (0.3 ± 0.3 ns from $z = 2.5$ nm to $z = 1.5$ nm).

Importantly, there is a discrepancy in the nature and extent of bilayer perturbation sampled during nonequilibrium unrestrained vs equilibrium restrained simulations. In Figure 6, we compare the local reorganization of the proximal bilayer leaflet that occurred during unrestrained and US simulations. During unrestrained sampling, the most populated bilayer state changed from a depression to a protrusion with deeper solute immersion near $z = 1.9$ nm (Figure 6). However, in the restrained simulations, this transition occurred near $z = 2.1$ nm (Figure 6).

To evaluate the possibility that this discrepancy results from systematic sampling errors at $1.9 \leq z \leq 2.1$ nm in our restrained simulations, we conducted additional US simulations at $z_i^0 = 2.0$ nm. These simulations directly probe the probability that US simulations may be trapped in metastable bilayer protrusion states at $1.9 \leq z \leq 2.1$ nm due to hidden sampling barriers along z_p^{upper} , the deviation from the mean of the distance along z between the center of the bilayer and the phosphorus atom of lipids in the upper leaflet. These simulations were initiated with conformations drawn from unrestrained simulations in which a methylpropane molecule located near $z = 2.0$ nm either did or did not subsequently progress into the hydrophobic interior of the bilayer. We refer to these starting conformations as binding and nonbinding, respectively. Not only do both binding (Figure 7A) and nonbinding (Figure 7B) initial conformations drawn from unrestrained simulations all predominantly sample bilayer protrusion states in US simulations at $z_i^0 = 2.0$ nm but they also do so to the same extent as the original set of US simulations at this depth (Figure 7C). Furthermore, rare transitions between local bilayer protrusion and depression are evident for a number of US simulations in Figure 7A and B, demonstrating that the apparent barrier along z_p^{upper} does not prevent the slow

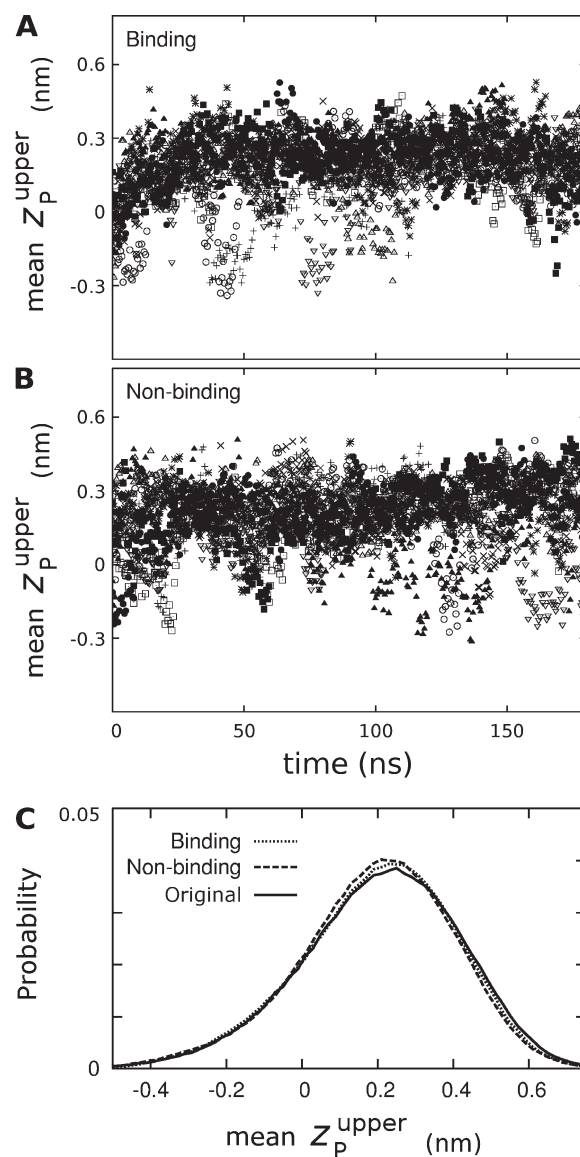


Figure 7. Bilayer reorganization as a function of simulation time from US at $z_i^0 = 2.0$ nm. (A and B) For each simulation, a different symbol represents the time series of the mean value of z_p^{upper} for lipids with $r_p < 1$ nm, block-averaged in 500 ps intervals. The reader is not expected to distinguish the different symbols throughout the sampling time. Rather, we use discrete symbols instead of depicting this data as a heat map to make it clear that excursions of sampling away from the main basin are transiently conducted by different simulations. Starting conformations were drawn from 10 snapshots obtained from unrestrained simulations in which the identified methylpropane molecule (A) subsequently bound the bilayer, progressing deeply into the hydrophobic interior, or (B) did not bind the bilayer, diffusing back into bulk water. (C) Histograms showing the probabilities at which values of z_p^{upper} were sampled during US initiated from (dotted line) productive binding snapshots, (dashed line) nonbinding snapshots, and (solid line) the 24 main US simulations of methylpropane in this work. The histogram bin width was 0.02 nm.

equilibration of protrusion and depression states at this depth on the simulation time scale. As discussed in the next section, this discrepancy questions the relevance of solute-restrained conformations to the binding mechanism.

DISCUSSION

Systematic Sampling Errors. The analysis of potential of mean-force free energy calculations for the insertion of small molecular solutes into a lipid bilayer shows that statistical convergence is limited by the rate of conformational reorganization of the lipid bilayer.

The accuracy of $\Delta G_{\text{bind}}^{\circ}$ can be strongly affected by systematic sampling errors occurring at specific locations along the reaction coordinate. In the previous study of methylpropane,²³ systematic sampling errors in only two umbrellas caused a systematic $\Delta G_{\text{bind}}^{\circ}$ error of 1 kcal/mol (Table 2). The two umbrellas that contained systematic sampling errors were centered at $z_i^0 = 1.9$ and 2.0 nm (Figure 2C). In a neat DOPC bilayer, this region corresponds to the rugged interface between the bilayer and water (Figure 3A). There, methylpropane is abruptly being dehydrated with increasing bilayer immersion, following a sharp transition from local depression to protrusion of the proximal bilayer–water interface with increasing solute depth between $z_i^0 = 2.1$ and 2.0 nm in the long simulations presented in this work (Figure 3B–D). In shorter simulations,²³ the two umbrella potentials centered at $z_i^0 = 1.9$ and 2.0 nm were trapped in a metastable state of leaflet depression similar to the one depicted in Figure 4A and did not sample the state of leaflet protrusion depicted in Figure 4B (data not shown), which is more favorable at these depths (Figure 3B–D). This finding is consistent with the significant spread in $\Delta G_{\text{bind}}^{\circ}$ values obtained from 20 ns samples per umbrella in this work (Figure 2A). Accordingly, the extension of the previously published simulations presented in this study resolved these systematic sampling errors (Table 2).

Systematic sampling errors were also revealed upon increasing the amount of sampling in simulations of *n*-propylguanidinium. In this case, the systematic $\Delta G_{\text{bind}}^{\circ}$ error is 3 kcal/mol (Table 2). In the previous study of *n*-propylguanidinium,²³ systematic sampling errors occurred near $z = 2.0$ and 2.4 nm (Figure 2F) and, similarly to the previous simulations of methylpropane,²³ also resulted from insufficient sampling of a state involving local lipid protrusion (Figures 3G–I). In this case, bilayer reorganization was concurrent with reorientation of the solute (Figure 3J), and the extension of the previously published simulations²³ for an additional 120 ns per umbrella did not resolve these systematic sampling errors, although the difference in $\Delta G_{\text{bind}}^{\circ}$ values was reduced (Table 2). Nevertheless, the 36 independent measures of $\Delta G_{\text{bind}}^{\circ}$ in this study range from -9.1 to -4.0 kcal/mol (Table 2), and it is only as an average that they converge to -6.7 ± 1.0 kcal/mol. The finding that a single set of US simulations lasting 205 ns per umbrella may be insufficient to obtain adequate convergence in the value of $\Delta G_{\text{bind}}^{\circ}$ for *n*-propylguanidinium should be regarded as an important cautionary tale for the calculation of equilibrium properties from simulations of peptides in lipid bilayers.

Increased sampling also affected the free energy profile of *n*-propylguanidinium near the center of the bilayer (Figure 2E,F). Here, changes in the orientation of the guanidino group (Figure 3J) are coupled to the presence, in one leaflet or the other, of a conic lipid surface defect that is lined by lipid headgroups (Figures 3F–I, 4F, 5F) and filled with water molecules that partially solvate the guanidino moiety (Figure 3F). Conformational states in which such defects occur in the distal leaflet are only metastable.⁵² The reason why *n*-propylguanidinium PMFs in this work show no indication of systematic sampling errors in the center of the bilayer for US

simulations with increasing t_{eq} is that the orientational auto-correlation time exceeds the simulation time scale (data not shown). That is, systematic sampling errors only become apparent in PMFs after a transition leading to a state of lower free energy occurs for the first time. Sampling errors in this region do not affect the computed binding free energy significantly because solute residence at the bilayer center is the least favorable state in the entire PMF (Figure 2E). However, the large kinetic barrier to convergence at the middle of the bilayer does affect the symmetry of the solute orientation about the bilayer center (Figure 3J) and the computed free energy barrier to bilayer traversal.⁵² At $z_i^0 = 0$ nm, both polarized orientations of the solute should be equally likely. Here, the presence of US restraining potentials, by preventing significant axial mobility of the solute, forces the system to cross pre-existing barriers that could otherwise be circumvented. This source of quasi-nonergodicity may be alleviated by the addition of equilibrium exchange.²²

Importantly, our results suggest that systematic sampling errors may also be present in other US studies of similar duration in which solutes interact with a lipid bilayer. Free energy profiles for the insertion of molecular solutes along the bilayer normal have been obtained using US and similar techniques for a long list of increasingly complex solutes including, among others, the small molecules hexane,¹³ oxygen, and ammonia;¹⁴ the steroids cholesterol⁵³ and cortisone;⁵⁴ the drugs valproic acid,¹¹ various adamantanes,¹² and the nonsteroidal anti-inflammatories acetylsalicylic acid and ibuprofen;¹⁰ a panel of amino acid side chain analogs;²³ and, finally, peptides including a model hexapeptide,⁵⁵ a transmembrane helix,⁵² the 18-residue cationic antimicrobial peptide protegrin-1,¹⁶ a 34-residue Kv channel gating-modifier toxin,⁵⁶ and a 42-residue fragment of the amyloid- β peptide.⁵⁷ In the aforementioned simulations, either $\Delta G_{\text{bind}}^{\circ}$ drifted systematically with increasing t_{eq} (see ref 55 and C.H. Davis, personal communication regarding ref 57), it depended significantly on the starting conformation,⁵⁶ or free-energy convergence measures based on t_{eq} were not conducted.^{10,11,13,14,16,23,32–54} From this list, only the study of adamantanes evaluated the systematic drift in the value of $\Delta G_{\text{bind}}^{\circ}$ based on t_{eq} and, using a statistical test, concluded that the simulations had converged after 15 ns per umbrella with $t_{\text{eq}} = 2$ ns.¹² In light of the present study, it remains possible that these simulations of adamantanes, amphipathic cationic molecules larger than *n*-propylguanidinium, contained systematic sampling errors but simply did not yet show it. Furthermore, it is important to keep in mind that some of the 36 PMFs that we constructed, each based on 205 ns of equilibration and sampling per umbrella, still contain unresolved systematic sampling errors. Finally, the simulations presented in this work may also be subject to systematic sampling errors that are not evident on this time scale.

Interestingly, the application of additional restraints on lipid headgroups appears to avoid the slow relaxation of $z_{\text{P}}^{\text{upper}}$ and speed up convergence of PMF calculations from US simulations of peptide insertion along the bilayer normal.⁵⁶ Note, however, that such restraints preclude large deformations of the lipid bilayer, which may be mechanistically important, such as those depicted in Figures 4 and 5.

Effect of Simulation Box Size. The simulations in this work employed a relatively small patch of bilayer, with only 32 lipids per leaflet. For the small solutes investigated in this study, systematic sampling errors persisted for up to 125 ns per umbrella (Figure 2D), and even longer equilibration times may be necessary for larger solutes. Thus, we chose to keep the

simulation box small in order to attain long simulation times. Importantly, at certain values of z_i^0 , the bilayer reorganization induced by the insertion of the solute involved the entire proximal leaflet (Figures 3B,G and 4B,F), suggesting that the finite size of the simulation box may result in a systematic bias of the PMF. This possible bias appears unlikely to dramatically affect the value of $\Delta G_{\text{bind}}^\circ$ for *n*-propylguanidinium. This is because bilayer reorganization remained predominantly local for $z_i^0 \geq 1.5$ nm (Figure 3G), which encompasses the region traversed by the solute from bulk water to the global minimum of the PMF at $1.4 \leq z \leq 1.5$ nm (Figure 2E). For methylpropane, however, bilayer reorganization induced by solute migration through the lipid–water interface involved most of the proximal leaflet for $2.1 \geq z_i^0 \geq 1.5$ nm (Figure 3B), a region of the PMF located between bulk water and the global minimum of the PMF at $z_i^0 = 0.0$ nm (Figure 2B), so that the box size may have a significant effect on the estimate of $\Delta G_{\text{bind}}^\circ$. A systematic evaluation of the dependence of $\Delta G_{\text{bind}}^\circ$ on bilayer size is an interesting avenue for future research. However, PMF calculations based on a conventional definition of solute insertion depth will be subjected to an increasingly degenerate reaction coordinate as the size of the bilayer is increased. This is because larger bilayers can experience larger undulations. When using very large bilayer patches, it may thus become possible for the COM of the solute to coincide with the COM of the bilayer along the bilayer normal even as the solute remains in bulk water.

Comparison to Experimental Transfer Free Energies. Our results indicate that previously published estimates of methylpropane and *n*-propylguanidinium binding free energies²³ contain systematic sampling errors. Nevertheless, those results²³ are in good agreement with the experimental water-to-cyclohexane transfer free energies reported by Radzicka and Wolfenden⁵⁸ for methylpropane (−4.9 kcal/mol) and *n*-propylguanidinium (+14.9 kcal/mol). This may be because the range of experimentally determined values for each side chain analog remains large in comparison to the theoretical uncertainties. For methylpropane, MacCallum et al. obtained a transfer free energy of -3.6 ± 0.4 kcal/mol,²³ using the value of the PMF at the bilayer center, whereas we obtain -4.8 ± 0.3 kcal/mol (Figure 2B). While the free energy value that we report appears to have improved accuracy, it should be noted that the experimental free energy quoted is actually the transfer free energy for *n*-butane.⁵⁸ In a separate simulation study, MacCallum and Tieleman used a thermodynamic cycle to calculate the free energy for the transfer of amino acid side chain analogs from water to cyclohexane.⁵⁹ The binding free energy value obtained for methylpropane using the OPLS-AA/L force field and a different treatment of long-range nonbonded interactions was, depending on the water model, −5.4 to −5.7 kcal/mol.⁵⁹ Alchemical pathways can be advantageously used to determine $\Delta G_{\text{bind}}^\circ$ in cases where the bound state is known. For methylpropane, an alchemical pathway can circumvent the hidden sampling barriers at the lipid–water interface. For *n*-propylguanidinium, however, and in the general case, knowledge of the PMF is required to define the bound state.

Moreover, it is difficult to make experimental comparisons for the value of $\Delta G_{\text{bind}}^\circ$ in the case of *n*-propylguanidinium because water-to-cyclohexane or -octanol transfer free energies are expected to differ from the free energy of transfer to the center or the interface of a DOPC bilayer, where the solute remains solvated by water molecules and lipid headgroups (Figure 3F) in a complex and highly anisotropic chemical environment (Figures 4D–F, 5D–F). In addition, the drastic reorganization

of the lipid bilayer that occurs in response to even small molecular solutes and the sensitivity of these perturbations to the nature of the solute (Figures 3–5) complicate the assumption of group additivity underlying the application of the free energy of partitioning of Ace-WLLxL peptides (where x includes L and R) between water and a bilayer similar to DOPC.⁶⁰ Indeed, the limitations of group contribution approaches have also been pointed out in simulation studies of hydration free energies.⁶¹

Relevance of Solute-Restrained Conformations to the Binding Mechanism. The bilayer reorganization induced by the insertion of methylpropane in equilibrium restrained simulations differs from the reorganization observed in nonequilibrium unrestrained simulations (Figure 6). Because conformations sampled by restrained simulations along physical reaction coordinates are attainable by unrestrained simulations, it is tempting to interpret the path of lowest free energy along z as the unrestrained solute binding pathway.⁵⁵ Generally, however, restrained simulations may not yield mechanistically representative states in the absence of a separation of time scales whereby unrestrained molecular diffusion along the reaction coordinate is much slower than diffusion along other degrees of freedom. This is because weak ergodicity breaking⁶² may occur during binding events in unrestrained simulations but not in restrained simulations over sufficiently long time scales (Figure 6). When spontaneous binding conformations near $z = 2.0$ nm are extracted from unrestrained simulations and used to initiate US simulations, the distribution of $z_{\text{p}}^{\text{upper}}$ requires 10 ns to converge to a state of local lipid protrusion (Figure 7A). However, during all 64 binding events observed in our unrestrained simulation, methylpropane passed through this region in significantly less than 1 ns. In other words, the conformation of the bilayer does not reach equilibrium as methylpropane freely diffuses inward through the head-group region. Schematically, when the reaction coordinate is not the slowest relaxing DOF, restrained simulations on sufficiently long time scales follow the path of lowest free energy in orthogonal DOFs (blue dotted line in Figure 1), whereas unrestrained sampling over the landscape presented in Figure 1 may most readily progress from α to β by way of γ if diffusion along the reaction coordinate (x axis) is faster than along other DOFs (y axis). This effect is similar to differences in experimental time averaging.⁶³

As a result, it would be incorrect to conclude from the time-averaged data in Figures 3B,C,D; 4B; and 5B that, during binding events, the bilayer is likely to protrude to encapsulate methylpropane when methylpropane is at a distance of 2.0 nm from the bilayer center. Indeed, Figure 6 demonstrates that, during spontaneous binding events, the bilayer interface is only likely to protrude once methylpropane is within 1.8 nm of the bilayer center. Inversely, it is likely that fluctuations within the bound state up to and including spontaneous unbinding events, none of which occurred in our unrestrained simulations, involve even larger protrusions of the bilayer surface than observed in the equilibrium picture obtained from ensemble averaging. Because it describes a reversible (i.e., infinitely slow) reaction, the free energy surface represents an ensemble average over both binding and unbinding events, regardless of the direction of travel of the solute and its history. In such a case, the lowest free-energy pathway through the saddle point in Figure 1 represents an average between trajectories following either γ or δ , depending on direction. This analysis underscores the importance of considering time-dependent information when one desires to draw mechanistic conclusions from equilibrium sampling.

General Implications to Membrane Solvation of Peptides and Proteins. The mechanisms of bilayer reorganization uncovered in the present study are likely to be relevant in the context of larger molecules such as peptides or proteins, in which a hydrophobic or charged moiety is anchored at a particular depth. In such cases, as seen here for *n*-propylguanidinium, the bound state may involve substantial deformation of the bilayer. In particular, the deformations outlined in Figures 4 and 5 are parts of bilayer adaptation to hydrophobic mismatches that may be introduced by transmembrane proteins.⁶⁴ During such adaptations, lipid molecules in close proximity to the solute adjust their own immersion depth to facilitate the inclusion of hydrophobic moieties and the exclusion of hydrophilic moieties from the hydrophobic region of the bilayer. Importantly, these phenomena may also exist in the interaction of lipid bilayers with interfacial membrane proteins or peptides.

Interestingly, the states depicted in Figures 4 and 5 suggest that peripheral membrane proteins or interfacial peptides could induce bilayer bending by inducing changes in lipid ordering. The induction of bilayer bending is important in any cellular process that relies on membrane tension, including endocytosis, exocytosis, cell motility, and viral membrane fusion during infection.⁶⁵ Furthermore, protein-induced changes in membrane tension are capable of feeding back into the function of mechanosensitive membrane proteins.⁶⁶

CONCLUSIONS

The above results suggest that the difficulty of attaining convergence in simulations of molecular solutes embedded in lipid bilayers has, in general, been drastically underestimated due to the presence of hidden sampling barriers involving the slow reorganization of the lipid–water interface in response to solute insertion. Although we focused our analysis of systematic sampling errors on a single study,²³ it seems likely that similar systematic sampling errors pervade the literature. It is unfortunate that many published simulation studies do not report measures of convergence, since this information contributes crucially to the inferred accuracy of the measured quantities.⁶⁷ The publication of convergence measures, especially those that are capable of detecting systematic drifts in values obtained with increasing equilibration time, should be systematically encouraged.

On the basis of the above analysis, we propose that simulation studies could benefit from the following suggestions:

1. Performing multiple distinct sets of simulations of the same system from different starting conformations and computing PMFs separately from each set. Degrees of freedom predicted to contribute significantly to the binding free energy should be evaluated systematically. It may also be useful to use a variety of methods to generate these starting conformations in order to vary the starting conformations in unexpected ways.
2. Evaluating a key observable(s), such as the binding free energy, as a function of both simulation time and equilibration time and reporting these convergence measures.

AUTHOR INFORMATION

Corresponding Author

*E-mail: pomes@sickkids.ca.

ACKNOWLEDGMENT

Computations were performed on the GPC supercomputer at the SciNet HPC Consortium.⁶⁸ SciNet is funded by the Canada

Foundation for Innovation under the auspices of Compute Canada; the Government of Ontario; Ontario Research Fund - Research Excellence; and the University of Toronto. C.N. is funded by the Research Training Center at the Hospital for Sick Children and by the University of Toronto. This work was funded in part by CIHR Operating Grant MOP-43998. W.F.D.B. is supported by Alberta Heritage Foundation for Medical Research (AHFMR), Killam, and NSERC studentships. Work in D.P.T.'s group is funded in part by NSERC. D.P.T. is an AHFMR Scientist. R.P. is a CRCP chair holder.

REFERENCES

- (1) Edidin, M. Lipids on the frontier: a century of cell-membrane bilayers. *Nat. Rev. Mol. Cell Biol.* **2003**, *4* (5), 414–418.
- (2) Dyall, S. D.; Brown, M. T.; Johnson, P. J. Ancient invasions: from endosymbionts to organelles. *Science* **2004**, *304* (5668), 253–257.
- (3) Deamer, D. W.; Bramhall, J. Permeability of lipid bilayers to water and ionic solutes. *Chem. Phys. Lipids* **1986**, *40* (2–4), 167–188.
- (4) Engelman, D. M. Membranes are more mosaic than fluid. *Nature* **2005**, *438* (7068), 578–580.
- (5) Simons, K.; Vaz, W. Model systems, lipid rafts, and cell membranes. *Annu. Rev. Biophys. Biomol. Struct.* **2004**, *33* (1), 269–295.
- (6) Egberts, E.; Berendsen, H. J. C. Molecular dynamics simulation of a smectic liquid crystal with atomic detail. *J. Chem. Phys.* **1988**, *89* (6), 3718–3732.
- (7) Tieleman, D. P.; Marrink, S. J.; Berendsen, H. J. C. A computer perspective of membranes: molecular dynamics studies of lipid bilayer systems. *BBA—Rev. Biomembranes* **1997**, *1331* (3), 235–270.
- (8) Takaoka, Y.; Pasenkiewicz-Gierula, M.; Miyagawa, H.; Kitamura, K.; Tamura, Y.; Kusumi, A. Molecular dynamics generation of nonarbitrary membrane models reveals lipid orientational correlations. *Biophys. J.* **2000**, *79* (6), 3118–3138.
- (9) Poger, D.; Mark, A. E. On the validation of molecular dynamics simulations of saturated and *cis*-monounsaturated phosphatidylcholine lipid bilayers: a comparison with experiment. *J. Chem. Theory Comput.* **2010**, *6* (1), 325–336.
- (10) Boggara, M. B.; Krishnamoorti, R. Partitioning of nonsteroidal antiinflammatory drugs in lipid membranes: a molecular dynamics simulation study. *Biophys. J.* **2010**, *98* (4), 586–595.
- (11) Ulander, J.; Haymet, A. D. J. Permeation across hydrated DPPC lipid bilayers: simulation of the titratable amphiphilic drug valproic acid. *Biophys. J.* **2003**, *85* (6), 3475–3484.
- (12) Chew, C. F.; Guy, A.; Biggin, P. C. Distribution and dynamics of adamantanes in a lipid bilayer. *Biophys. J.* **2008**, *95* (12), 5627–5636.
- (13) MacCallum, J. L.; Tieleman, D. P. Computer simulation of the distribution of hexane in a lipid bilayer: spatially resolved free energy, entropy, and enthalpy profiles. *J. Am. Chem. Soc.* **2006**, *128* (1), 125–130.
- (14) Marrink, S. J.; Berendsen, H. J. C. Permeation process of small molecules across lipid membranes studied by molecular dynamics simulations. *J. Phys. Chem.* **1996**, *100* (41), 16729–16738.
- (15) Bemporad, D.; Luttmann, C.; Essex, J. W. Computer simulation of small molecule permeation across a lipid bilayer: dependence on bilayer properties and solute volume, size, and cross-sectional area. *Biophys. J.* **2004**, *87* (1), 1–13.
- (16) Vivcharuk, V.; Kaznessis, Y. Free energy profile of the interaction between a monomer or a dimer of protegrin-1 in a specific binding orientation and a model lipid bilayer. *J. Phys. Chem. B* **2010**, *114* (8), 2790–2797.
- (17) Yeh, I.-C.; Olson, M. A.; Lee, M. S.; Wallqvist, A. Free-energy profiles of membrane insertion of the M2 transmembrane peptide from influenza A virus. *Biophys. J.* **2008**, *95* (11), 5021–5029.
- (18) van Gunsteren, W. F.; Dolenc, J.; Mark, A. E. Molecular simulation as an aid to experimentalists. *Curr. Opin. Struct. Biol.* **2008**, *18* (2), 149–153.
- (19) Zuckerman, D. M. Equilibrium sampling in biomolecular simulations. *Ann. Rev. Biophys.* **2011**, *40* (1), 41–62.

- (20) Torrie, G. M.; Valleau, J. P. Nonphysical sampling distributions in Monte Carlo free-energy estimation: umbrella sampling. *J. Comput. Phys.* **1977**, *23* (2), 187–199.
- (21) Roux, B. The calculation of the potential of mean force using computer simulations. *Comput. Phys. Commun.* **1995**, *91* (1–3), 275–282.
- (22) Neale, C.; Roderger, T.; Pomès, R. Equilibrium exchange enhances the convergence rate of umbrella sampling. *Chem. Phys. Lett.* **2008**, *460* (1–3), 375–381.
- (23) MacCallum, J. L.; Bennett, W. F. D.; Tieleman, D. P. Distribution of amino acids in a lipid bilayer from computer simulations. *Biophys. J.* **2008**, *94* (9), 3393–3404.
- (24) Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. GROMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theory Comput.* **2008**, *4* (3), 435–447.
- (25) Jorgensen, W.; Chandrasekhar, J.; Madura, J.; Impey, R.; Klein, M. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79* (2), 926–935.
- (26) 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* (45), 11225–11236.
- (27) Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L. Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. *J. Phys. Chem. B* **2001**, *105* (28), 6474–6487.
- (28) Berger, O.; Edholm, O.; Jähnig, F. Molecular dynamics simulations of a fluid bilayer of dipalmitoylphosphatidylcholine at full hydration, constant pressure, and constant temperature. *Biophys. J.* **1997**, *72* (5), 2002–2013.
- (29) Chakrabarti, N.; Neale, C.; Payandeh, J.; Pai, E. F.; Pomès, R. An iris-like mechanism of pore dilation in the CorA magnesium transport system. *Biophys. J.* **2010**, *98* (5), 784–792.
- (30) Darden, T.; York, D.; Pedersen, L. Particle mesh Ewald: An $N \cdot \log(N)$ method for Ewald sums in large systems. *J. Chem. Phys.* **1993**, *98* (12), 10089–10092.
- (31) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. A smooth particle mesh Ewald method. *J. Chem. Phys.* **1995**, *103* (19), 8577–8593.
- (32) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **1984**, *81* (8), 3684–3690.
- (33) Miyamoto, S.; Kollman, P. A. Settle: An analytical version of the SHAKE and RATTLE algorithm for rigid water models. *J. Comput. Chem.* **1992**, *13* (8), 952–962.
- (34) Hess, B. P-LINCS: a parallel linear constraint solver for molecular simulation. *J. Chem. Theory Comput.* **2008**, *4* (1), 116–122.
- (35) Kandt, C.; Ash, W. L.; Tieleman, D. P. Setting up and running molecular dynamics simulations of membrane proteins. *Methods* **2007**, *41* (4), 475–488.
- (36) Hristova, K.; Wimley, W. C. A look at arginine in membranes. *J. Membr. Biol.* **2011**, *239* (1–2), 49–56.
- (37) Roux, B. Lonely arginine seeks friendly environment. *J. Gen. Physiol.* **2007**, *130* (2), 233–236.
- (38) Yoo, J.; Cui, Q. Does arginine remain protonated in the lipid membrane? Insights from microscopic pKa calculations. *Biophys. J.* **2008**, *94* (8), L61–L63.
- (39) MacCallum, J. L.; Bennett, W. F. D.; Tieleman, D. P. Partitioning of amino acid side chains into lipid bilayers: results from computer simulations and comparison to experiment. *J. Gen. Physiol.* **2007**, *129* (5), 371–377.
- (40) Li, L.; Vorobyov, I.; MacKerell, A. D.; Allen, T. W. Is arginine charged in a membrane?. *Biophys. J.* **2008**, *94* (2), L11–L13.
- (41) Lindahl, E.; Hess, B.; van der Spoel, D. GROMACS 3.0: a package for molecular simulation and trajectory analysis. *J. Mol. Model.* **2001**, *7* (8), 306–317.
- (42) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Hermans, J. Interaction models for water in relation to protein hydration. In *Intermolecular Forces*; Pullman, B., Ed.; D. Reidel Publishing Company: Dordrecht, The Netherlands, 1981; pp 331–342.
- (43) Grossfield, A. WHAM: the weighted histogram analysis method. <http://membrane.urmc.rochester.edu/content/wham> (accessed August 3, 2011).
- (44) Kumar, S.; Rosenberg, J. M.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A. The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *J. Comput. Chem.* **1992**, *13* (8), 1011–1021.
- (45) Ben-Tal, N.; Ben-Shaul, A.; Nicholls, A.; Honig, B. Free-energy determinants of alpha-helix insertion into lipid bilayers. *Biophys. J.* **1996**, *70* (4), 1803–1812.
- (46) General, I. J. A note on the standard state's binding free energy. *J. Chem. Theory Comput.* **2010**, *6* (8), 2520–2524.
- (47) Grossfield, A.; Woolf, T. B. Interaction of tryptophan analogs with POPC lipid bilayers investigated by molecular dynamics calculations. *Langmuir* **2002**, *18* (1), 198–210.
- (48) Humphrey, W.; Dalke, A.; Schulten, K. VMD: visual molecular dynamics. *J. Mol. Graphics* **1996**, *14* (1), 33–38.
- (49) Flyvbjerg, H.; Petersen, H. G. Error estimates on averages of correlated data. *J. Chem. Phys.* **1989**, *91* (1), 461–466.
- (50) Bogusz, S.; Cheatham, T. E.; Brooks, B. R. Removal of pressure and free energy artifacts in charged periodic systems via net charge corrections to the Ewald potential. *J. Chem. Phys.* **1998**, *108* (17), 7070–7084.
- (51) Berendsen, H. J. C.; Marrink, S.-J. Molecular dynamics of water transport through membranes: water from solvent to solute. *Pure Appl. Chem.* **1993**, *65* (12), 2513–2520.
- (52) Li, L.; Vorobyov, I.; Allen, T. Potential of mean force and pKa profile calculation for a lipid membrane-exposed arginine side chain. *J. Phys. Chem. B* **2008**, *112* (32), 9574–9587.
- (53) Zhang, Z.; Lu, L.; Berkowitz, M. L. Energetics of cholesterol transfer between lipid bilayers. *J. Phys. Chem. B* **2008**, *112* (12), 3807–3811.
- (54) Vijayan, R.; Biggin, P. C. A steroid in a lipid bilayer: localization, orientation, and energetics. *Biophys. J.* **2008**, *95* (7), L45–L47.
- (55) Babakhani, A.; Gorfe, A. A.; Kim, J. E.; McCammon, J. A. Thermodynamics of peptide insertion and aggregation in a lipid bilayer. *J. Phys. Chem. B* **2008**, *112* (34), 10528–10534.
- (56) Wee, C. L.; Gavaghan, D.; Sansom, M. S. P. Lipid bilayer deformation and the free energy of interaction of a Kv channel gating-modifier toxin. *Biophys. J.* **2008**, *95* (8), 3816–3826.
- (57) Davis, C. H.; Berkowitz, M. L. Interaction between amyloid- β (1–42) peptide and phospholipid bilayers: a molecular dynamics study. *Biophys. J.* **2009**, *96* (3), 785–797.
- (58) Radzicka, A.; Wolfenden, R. Comparing the polarities of the amino acids: side-chain distribution coefficients between the vapor phase, cyclohexane, 1-octanol, and neutral aqueous solution. *Biochemistry* **1988**, *27* (5), 1664–1670.
- (59) MacCallum, J. L.; Tieleman, D. P. Calculation of the water–cyclohexane transfer free energies of neutral amino acid side-chain analogs using the OPLS all-atom force field. *J. Comput. Chem.* **2003**, *24* (15), 1930–1935.
- (60) Wimley, W. C.; White, S. H. Experimentally determined hydrophobicity scale for proteins at membrane interfaces. *Nat. Struct. Biol.* **1996**, *3* (10), 842–848.
- (61) König, G.; Boresch, S. Hydration free energies of amino acids: why side chain analog data are not enough. *J. Phys. Chem. B* **2009**, *113* (26), 8967–8974.
- (62) Bouchaud, J. P. Weak ergodicity breaking and aging in disordered systems. *J. Phys. I France* **1992**, *2* (9), 1705–1713.
- (63) Paula, L. C.; Wang, J.; Leite, V. B. P. Statistics and kinetics of single-molecule electron transfer dynamics in complex environments: a simulation model study. *J. Chem. Phys.* **2008**, *129* (22), 224504.
- (64) Kandasamy, S. K.; Larson, R. G. Molecular dynamics simulations of model trans-membrane peptides in lipid bilayers: a systematic investigation of hydrophobic mismatch. *Biophys. J.* **2006**, *90* (7), 2326–2343.
- (65) Nambiar, R.; McConnell, R. E.; Tyska, M. J. Control of cell membrane tension by myosin-I. *P. Natl. Acad. Sci. U. S. A.* **2009**, *106* (29), 11972–11977.

(66) Phillips, R.; Ursell, T.; Wiggins, P.; Sens, P. Emerging roles for lipids in shaping membrane-protein function. *Nature* **2009**, *459* (7245), 379–385.

(67) Lyman, E.; Zuckerman, D. M. Ensemble-based convergence analysis of biomolecular trajectories. *Biophys. J.* **2006**, *91* (1), 164–172.

(68) Loken, C.; Gruner, D.; Groer, L.; Peltier, R.; Bunn, N.; Craig, M.; Henriques, T.; Dempsey, J.; Yu, C.-H.; Chen, J.; Dursi, L. J.; Chong, J.; Northrup, S.; Pinto, J.; Knecht, N.; Van Zon, R. SciNet: lessons learned from building a power-efficient top-20 system and data centre. *J. Phys.: Conf. Ser.* **2010**, *256* (1), 012026.

Determination of Conformational Equilibria in Proteins Using Residual Dipolar Couplings

Alfonso De Simone,^{†,‡} Rinaldo W. Montalvao,[†] and Michele Vendruscolo^{*,†}

[†]Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, United Kingdom

[‡]Division of Molecular Biosciences, Imperial College, South Kensington Campus, London SW7 2AZ, United Kingdom

 Supporting Information

ABSTRACT: In order to carry out their functions, proteins often undergo significant conformational fluctuations that enable them to interact with their partners. The accurate characterization of these motions is key in order to understand the mechanisms by which macromolecular recognition events take place. Nuclear magnetic resonance spectroscopy offers a variety of powerful methods to achieve this result. We discuss a method of using residual dipolar couplings as replica-averaged restraints in molecular dynamics simulations to determine large amplitude motions of proteins, including those involved in the conformational equilibria that are established through interconversions between different states. By applying this method to ribonuclease A, we show that it enables one to characterize the ample fluctuations in interdomain orientations expected to play an important functional role.

INTRODUCTION

The accurate determination of the conformational transitions associated with the interactions between proteins and their partners is a key challenge to understand the mechanism of action of these macromolecules.^{1–4} A view that is rapidly gaining momentum is that proteins can recognize and bind their partners because in their free states they often already transiently populate the conformations that they adopt in the bound states.^{2,3,5,6} A major role in the establishment of this idea has been played by nuclear magnetic resonance (NMR) spectroscopy, which is a technique that can provide detailed information about the dynamics of proteins.^{2,7–11} Residual dipolar couplings (RDCs), which report on the orientation of interatomic bonds with respect to an external magnetic field,^{12,13} are most useful in this context. A number of procedures to employ RDCs for the characterization of the structure and dynamics of proteins and nucleic acids have been proposed, including analytical deconvolution,^{14–16} the Gaussian axial fluctuations method,^{10,17} restrained molecular dynamics simulations in which the alignment tensor is either fitted to the experimental RDCs^{18–22} or calculated directly from the structure,^{23,24} and direct comparison with molecular dynamics simulations.^{10,25–29}

We discuss here an approach in which RDCs are used to determine ensembles of structures representing conformational equilibria of proteins undergoing large-amplitude motions. In this approach, RDCs are employed as replica-averaged restraints in molecular dynamics simulations to bias the trajectory of the protein molecules toward the state observed experimentally. In order to reproduce the time and ensemble averaging that is implicit in NMR measurements, several molecules (the replicas) are simulated simultaneously, and NMR parameters are back-calculated as average values over them. To implement the restraints, RDCs are calculated from the shape and charge of each individual replica in the set used in the simulations,^{25,30–34}

and the resulting average values are then required to match the measured RDCs. Since in heterogeneous states of proteins and nucleic acids the alignment tensors, as well as the corresponding RDCs of the individual conformations populated during the dynamics, can differ very significantly,^{29,30,35,36} this approach, which does not require the assumption that the fluctuations of the alignment tensor remain relatively small and that they are uncorrelated with the fluctuations of the structure,^{29,35,36} is expected to be particularly suitable for enabling the description of the conformational interconversions associated with the function of these macromolecules.

In this work we present a validation of this method by considering the test of the “reference ensemble”.^{22,37,38} In this test a reference ensemble is generated at first by using molecular dynamics simulations with a given force field. A set of RDC values is then back-calculated from the structures of this ensemble. These RDC values, which are not measured experimentally but derived computationally from known structures, are then used as restraints in new molecular dynamics simulations with another force field. In the absence of the RDC restraints, this second force field gives rise to an ensemble of conformations different from the reference ensemble. The presence of the RDC restraints, however, induces the second force field to sample an ensemble of conformations closely reproducing the starting reference ensemble, provided that the restraints are implemented correctly. Since in this procedure one knows exactly the ensemble of structures that give rise to the RDC values used as restraints, it is possible to assess very accurately whether the restraints are effectively implemented to bias the conformational sampling toward the reference ensemble. We show that by applying this method, it is possible to characterize motions of large amplitude in ribonuclease A (RNase A). RNase A

Received: May 30, 2011

Published: October 10, 2011

is a V-shaped protein whose concerted motions of the two antiparallel β -sheet regions (V1 and V2) are closely connected to its function, which involves substrate binding and release.^{39–41} These low-frequency “breathing” motions, which have been extensively studied experimentally,^{41–43} make RNase A a paradigmatic system to study the relationship between conformational fluctuations and function.

RESULTS

Test of the Reference Ensemble. In order to carry out a rigorous test of the method for carrying out molecular dynamics simulations with replica-averaged RDC restraints that we discuss in this work, we applied the test of the reference ensemble.^{22,37,38} For this method, a ‘reference ensemble’ of conformations was generated by unrestrained molecular dynamics simulations (see Reference Ensemble Calculations in the Materials and Methods Section). RDCs were calculated from the structures of this reference ensemble (see Structure-Based Calculation of the Alignment Tensors in the Materials and Methods Section) and employed as structural restraints (see Restrained Ensemble Calculations in the Materials and Methods Section) to generate a ‘restrained ensemble’, which was then shown to closely reproduce the conformational properties of the original reference ensemble. The advantage of using this reference ensemble test is that it allows for a stringent validation analysis in which the atomic coordinates of the conformations in the reference ensemble are known exactly, and therefore, the accuracy of the conformations in the restrained ensemble can be assessed with great confidence.^{22,37,38}

Generation of the Reference Ensemble. The reference ensemble was generated by a 100 ns unrestrained trajectory of RNase A by using the Gromos96 force field; for comparison an ‘unrestrained ensemble’ was generated from a 100 ns trajectory by using the Amber99SB force field (see Reference Ensemble Calculations in the Materials and Methods Section). The two force fields generated two different types of breathing motions in the native state of RNase A, with Amber99SB oscillating around a moderately open state and Gromos96 populating both open and closed conformations (Figure S1, Supporting Information). The differences between the reference and unrestrained ensembles arise in part from the fact that the Gromos96 force fields implements a united-atom representation, which speeds up the sampling of the conformational space.

Structure-Based Alignment Approach in Molecular Dynamics Simulations. To enforce the RDCs as structural restraints in molecular dynamics simulations, we implemented our own version of the PALES method^{30,34} for both steric and electrostatic alignment media in the GROMACS package (see Structure-Based Calculation of the Alignment Tensors in the Materials and Methods Section). The employment of structure-based calculations of the alignment tensor during the sampling is crucial to obtain an accurate determination of conformationally heterogeneous states of proteins, as the alignment that these molecules can adopt can vary significantly during the dynamics (Figure 1), and it is therefore often not possible to make the approximation that all the structures in the ensemble have the same alignment tensor.^{29,35,36}

Replica-Averaged RDC Restraints. The approach that we followed in this study includes an averaging of the RDCs over multiple replicas of the protein molecule, which allows experimental

restraints to be imposed as an ensemble property.^{18,19,22,44} This approach is particularly efficient in sampling protein ensembles representing conformationally heterogeneous states because the replica averaging allows populating simultaneously different conformational basins that are present in solution and that contribute to the experimental observables.^{19,44,45} In this paper, by analyzing the effects of using different numbers of alignment media and different numbers of replicas (Figures S2 and S3, Supporting Information), we defined an optimal protocol based on replica-averaged RDC restraints with 16 replicas and RDCs data for two bond vectors from three alignment media. The employment of *in silico* experiments allows also for accurately assessing the effects of the errors on the RDC data in the performance of the structure-based alignment prediction (Figure S4, Supporting Information).

RDC-Based Determination of Large-Scale Motions of RNase A. The main result of this work is the demonstration that it is possible to use information derived from RDCs to characterize with high-accuracy distinct states in conformational equilibrium. To illustrate our approach, we considered the open and the closed forms of RNase A. In order to follow the conformational changes involved in the breathing motions of RNase A, we used the angle (‘pincer angle’) between the centers of mass of V1 and V2 and their hinge region (Figure 2a), which discerns between the open and closed conformations. We show that, while only the open state is present in the unrestrained Amber99SB simulations (Figure 2b), the use of RDC restraints drives a series of conformational interconversions between the open and closed states, which generates an ensemble with essentially correct Boltzmann weights despite the presence of an underlying force field that has a different behavior when considered on its own (Figure 2d). We further validate these results by reporting the Q factors (see Materials and Methods) for the unrestrained and restrained ensembles compared to the reference ensembles. In addition to the bonds vectors employed for generating the restrained ensemble, which expectedly exhibited a better agreement in the case of the restrained ensemble, all other bond vectors analyzed showed improved Q factors for the restrained ensemble (Figure 2c).

Structural Accuracy of the RDC-Driven Sampling. In order to investigate more in detail how structure-based calculations of the alignment tensor can be used to drive conformational transitions, we calculated the RDCs corresponding to a selected closed conformation from the reference ensemble and used them to restrain one-replica molecular dynamics simulations that started from the open state and ended up in the closed state (Figure 2a). These simulations show that by imposing the RDC restraints on an open conformation it is possible in 1 ns to drive the conformational sampling toward the pincer angle corresponding to the closed conformation (Figure 2a), with a very significant reduction of the root-mean-square deviation (rmsd) from the closed conformation itself (from ~ 5.3 to ~ 0.5 Å) (Figures S5 and 6S, Supporting Information). For comparison, by using an alternative approach in which the conformations that better reproduce the experimental data are selected from an unrestrained sampling, we could only obtain a relatively poor agreement with RDCs ($Q = 0.43$, Figure 3b) compared to the restrained simulations ($Q = 0.16$). We found the conformations selected from the unrestrained ensemble to be all quite far from the target structure in terms of pincer angle values (Figure 3a, red line), thereby resulting in an inaccurate conformational ensemble.

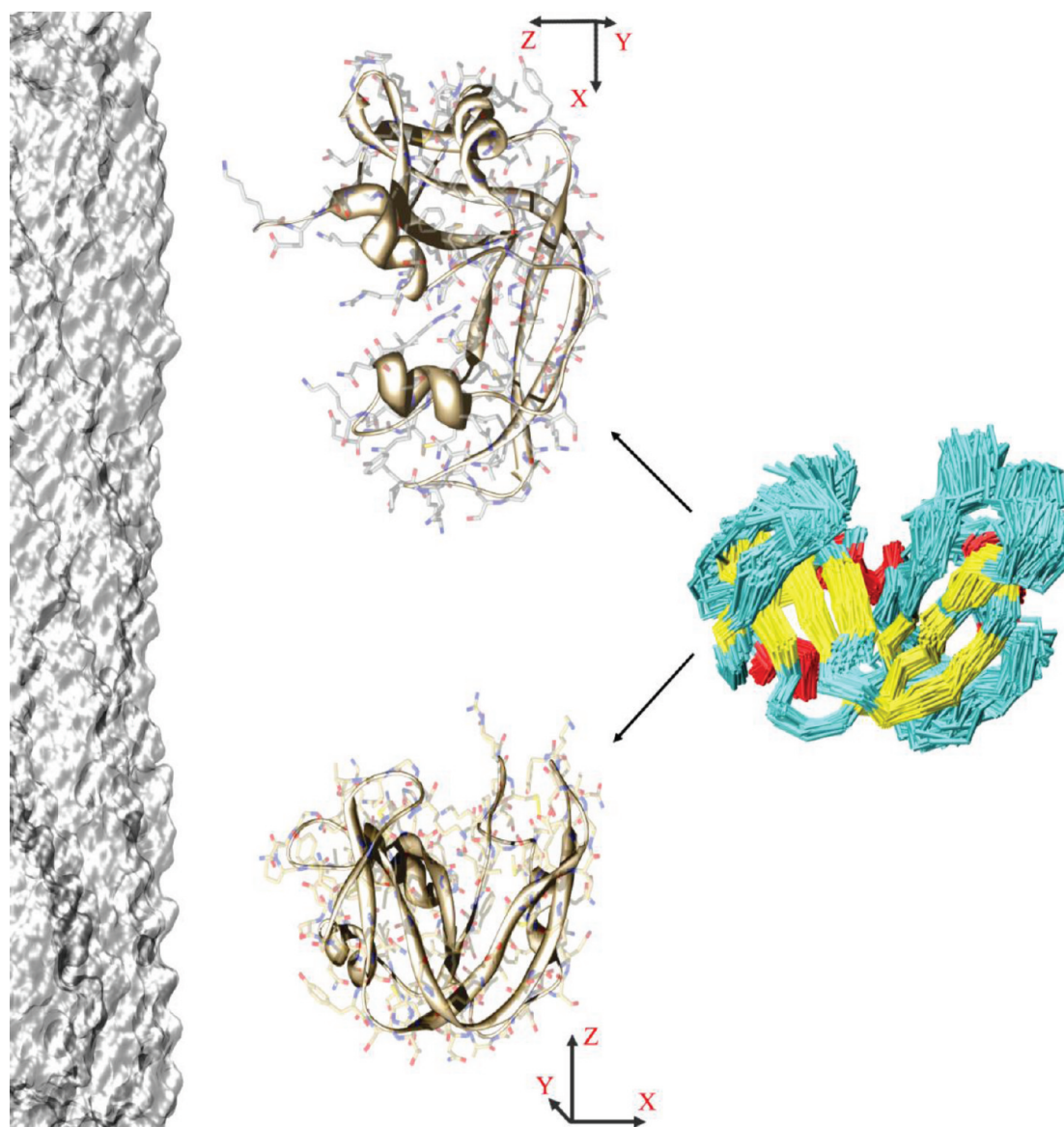


Figure 1. The alignment tensor is highly sensitive to the conformation of the protein. The preferential orientations of a protein with respect an alignment medium, especially in electrostatic cases, can vary significantly within the conformational ensemble of the protein. Such dependence is illustrated here by considering two conformations within the RNase A reference ensemble that have very different alignment tensors in the presence of Pfl (shown in surface representation on the left).

DISCUSSION

Our knowledge of the three-dimensional states that biological macromolecules adopt in solution has enormously improved in recent years.^{8,9,46–49} It has also been established that even in their native states, proteins constantly undergo structural fluctuations with time scales ranging from picoseconds to seconds and beyond, which are biologically relevant and influence a wide variety of processes, including enzymatic catalysis, ligand binding, and the formation of biomolecular complexes.^{2,7–11,44} States of this type pose a formidable challenge for structure determination, because, in many cases, they are inherently flexible and conformationally highly heterogeneous.

We have described a computational procedure for using NMR measurements of residual dipolar couplings as replica-averaged restraints in molecular dynamics simulations to determine large amplitude motions of proteins, including those involved in the

conformational equilibria often associated to their functions. When proteins undergo significant motions, as in the case of RNase A considered here, the method that we have described is highly effective in exploiting the structural and dynamical information provided by RDCs to determine accurately conformational ensembles and the associated Boltzmann weights (Figure 2d). For comparison, related methods for determining the alignment tensor by a fit to experimental RDCs^{18–22} can be expected to be accurate primarily when the conformational heterogeneity of the solution state is limited; this is because the experimental RDCs, which are averaged over the molecules in solution, can be significantly different from the specific RDCs of individual conformations in the corresponding ensemble, and their alignment tensors can be very different from each other and depart substantially from the average alignment tensor (Figures 1 and 3c). It should also be noted that, when the conformational

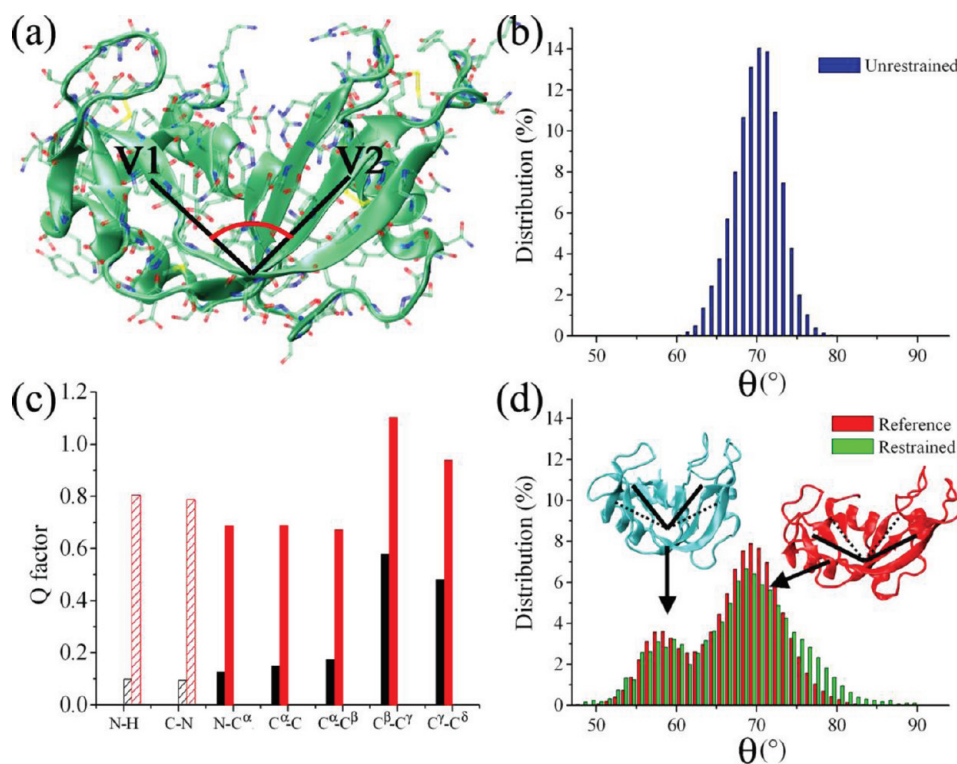


Figure 2. Determination of large-amplitude structural fluctuations from RDCs. (a) Representation of the structure of RNase A (Protein Data Bank code 7RSA [32]); the pincer angle, which accounts for the large motions between the antiparallel β -sheets V1 (residues 61–63, 71–75, 105–111, and 116–12) and V2 (residues 42–46, 82–87, and 96–101), is indicated schematically. The value of the pincer angle is calculated from the three centers of mass of the $C\alpha$ -atoms of three protein regions: region 1 (V2) spanning residues 42 and 43, region 2 (hinge) spanning residues 48, 49, and 80, and region 3 (V1) spanning residues 72 and 73. (b) Pincer angle distribution in 100 ns unrestrained Amber99SB simulations. (c) Agreement (Q factor) between RDCs of the reference and restrained ensembles (black) and the reference and unrestrained ensembles (red); dashed lines indicate bond vectors employed as restraints. (d) Pincer angle distribution in the reference (red) and restrained (green) ensembles calculated with three alignment media and 16 replicas (Figure S6, Supporting Information); the bimodal distribution includes both closed (blue) and open (red) conformations. Thirty structures per replica are recorded in the final part of each cycle (sampled at 300 K) of the restrained ensemble with a total of 9600 conformations.

fluctuations are of large amplitude, the structural interpretation of RDCs that is adopted here may raise concerns, especially when the time scale of alignment is faster than that of interconversion between different conformers.²³ Since, however, RDCs are measured under equilibrium conditions, the specific features of the dynamics of the alignment process should not affect their values. Another concern derives from the possibility of perturbing the conformational properties through the process of measurement itself. Also in this case, however, as long as the interaction between proteins and alignment media is weak so that only the orientation of the proteins with respect to an external reference frame, but not their internal conformational space, is altered significantly by the presence of the alignment media. We suggest that for these reasons the structure-based calculation of the alignment tensors implemented here to enforce the replica-averaged RDC restraints can provide an accurate representation of conformational equilibria.

In conclusion, we have described a method of using RDCs as replica-averaged restraints in molecular dynamics simulations to provide a quantitative description of the free energy landscapes associated with large-scale motions of proteins.

MATERIAL AND METHODS

Reference Ensemble Calculations. Unrestrained molecular dynamics simulations were performed with the Gromacs package.⁵⁰

Two independent molecular dynamics simulations have been carried out. In the first simulation we generated the ‘unrestrained ensemble’, employing the all-atom Amber99SB force field^{51,52} and the TIP3P explicit water model,⁵³ and in the second simulation we generated the ‘reference ensemble’, employing the Gromos96 force field⁵⁴ and the SPCE water model.⁵⁵ The starting coordinates were derived from the crystal structure of the RNase A (Protein Data Bank code 7RSA).⁵⁶ All simulations were carried out in the NPT ensemble with periodic boundary conditions at a constant temperature of 300 K and pressure of 1 atm. A dodecahedron box was employed for accommodating the protein, water molecules, and ions. Bonds were constrained by the LINCS⁵⁷ algorithm. The particle-mesh Ewald (PME) method⁵⁸ was used to account for the electrostatic contribution to nonbonded interactions (grid spacing of 0.12 nm). To model a system at pH of 7, the protonation states of pH-sensitive residues were set as follows: Arg and Lys residues were positively charged, Asp and Glu residues were negatively charged, and His residues were neutral. The protonation state of His residues was derived by comparing high-resolution X-ray structures performed at different pH values;⁵⁹ His12 and His48 were protonated at N δ , and His105 and His119 were protonated at N ϵ . The net charge of the protein was neutralized by the addition of Cl⁻ ions. Simulations were continued for 100 ns. The Gromos96 trajectory was selected to provide the reference ensemble, whereas the

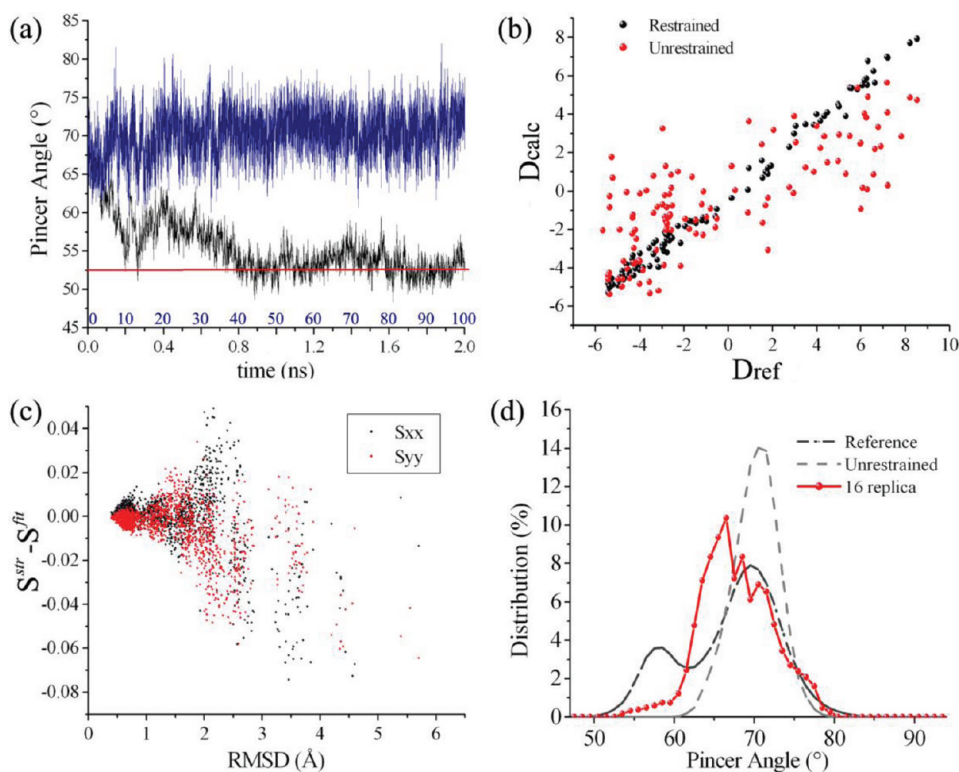


Figure 3. Structural accuracy of the RDC-driven sampling. (a) Time series of the pincer angle in the 100 ns unrestrained (Amber99SB) simulations (blue line) and 2 ns restrained (Amber99SB) simulations (black line), enforcing RDCs calculated from a closed structure of RNase A; the restraint force is gradually enforced during the first 1 ns. The red line indicates the angle value of the target structure. (b) Comparison between the RDCs of the target closed structure (D^{ref}), of the structures obtained in the last 1 ns of the restrained simulations (D^{calc} , black dots), and of the structures selected from the unrestrained simulations (D^{calc} , red dots). (c) Principal elements of the alignment tensors as a function of the rmsd between conformations extracted from the restrained simulation and the target structure from which the RDCs have been calculated; the plot shows the differences between fitted (S^{fit}) and structure-based alignment (S^{str}) tensors. (d) Comparison of the distribution of the pincer angle in the restrained ensemble obtained using only one steric alignment medium and 16 replicas (red) with the distributions from the reference (black dot-dashed) and unrestrained (gray dashed) ensembles.

Amber99SB trajectory was analyzed as an unrestrained ensemble to be compared with the restrained ensemble, which employed the same molecular dynamics settings with the addition of the RDC restraints (see Restrained Ensemble Calculations Section).

Restrained Ensemble Calculations. Molecular dynamics simulations with replica-averaged RDC restraints were implemented in the Gromacs package, by adopting a structure-based calculation of the alignment tensor (see Fitting Procedure for the Calculation of the Alignment Tensors Section). In this approach, restraints are imposed by adding a pseudoenergy term (E_{RDC}) to a standard molecular mechanics force field (E_{MM}):

$$E_{\text{TOT}} = E_{\text{MM}} + E_{\text{RDC}} \quad (1)$$

The resulting force field (E_{TOT}) was used in the molecular dynamics simulations. The E_{MM} that we employed was the Amber99SB (settings as in the unrestrained simulation section), and the pseudoenergy term is given by^{18–20,22}

$$E_{\text{RDC}} = \alpha \sum_i (D_i^{\text{exp}} - D_i^{\text{calc}})^2 \quad (2)$$

An initial equilibration simulation at 300 K was run, during which the agreement between calculated and experimental data, represented by their mean square deviation, eq 2, was allowed to converge. This result was achieved by gently raising the restraint force constant α . Subsequently, a series of 20 cycles of simulated annealing between 300 and 500 K was carried out to sample

effectively the conformational space. Each cycle was carried out for a total of 250 ps (125 000 molecular dynamics steps) by using an integration step of 2 fs. The restraints were imposed as averages over M replicas of the protein molecule; we employed simulations with $M = 2, 4, 8,$ and 16 . For each replica, the alignment tensors are independently computed using a structure-based method (see Structure-Based Calculation of the Alignment Tensors Section) and used in eq 2.

We are planning to include the implementation of the RDC restraints presented in this work in the standard Gromacs distribution.

Fitting Procedure for the Calculation of the Alignment Tensors. When we did not carry out structure-based calculations of the alignment tensor, the PALES code was used to perform calculations of the alignment tensor by fitting the RDCs to the structures.^{30,34}

Structure-Based Calculation of the Alignment Tensors. In order to calculate the RDCs corresponding to a given structure, we implemented our own version of the PALES code^{30,34} into the Gromacs package. We used three alignment media (Table 1): two steric (DMPC/DHPC and Pf1 at high ionic strength, Pf1-s) and an electrostatic (Pf1-e).

Definition of Q-Factors for RDCs. The quality factor for a given RDC (Q -factor) is defined as⁶⁰

$$Q_i = \sqrt{\frac{(D_i^{\text{ref}} - D_i^{\text{res}})^2}{(D_i^{\text{ref}})^2}}$$

Table 1. Comparison of the Three Alignment Media Used in This Work^a

	NSPE	correlation	SD (Hz)	Q-factor
Pf1-e vs Pf1-s	-0.32	0.45	11.2	1.375
Pf1-e vs DHCP/DMCP	0.3	0.31	15.67	1.219
Pf1-s vs DHCP/DMCP	-0.54	0.86	8.62	0.533

^aThree alignment media: two steric (DMCP/DHCP and Pf1 at high ionic strength, Pf1-s) and an electrostatic (Pf1 at low ionic strength, Pf1-e).

where D_i^{ref} is the RDC of a bond i in the reference ensemble and D_i^{res} is the corresponding RDC in the restrained ensemble.

Comparison of the Alignment Tensors. The independence of the RDC sets was tested by four similarity indexes (Table 1). In addition to standard analyses (i.e., correlation, standard deviation, Q-factor), we introduced the normalized scalar product of alignment tensors eigenvectors (NSPE):

$$\text{NSPE}_{ab} = \frac{1}{3} \sum_{i=1}^3 (\vec{d}_a(i) \cdot \vec{d}_b(i))$$

where $\vec{d}_a(i)$ and $\vec{d}_b(i)$ represent the three eigenvectors of the alignment tensors of the media a and b , respectively. The NSPE index ranges from -1 , for completely opposite alignment tensors, to 1 for totally overlapping tensors.

RDCs Calculation from the Reference Ensemble. For a given protein structure, the RDC calculated on the bond vector between atoms P and Q is given by

$$D_{\text{PQ}}^{\text{Calc}} = -\frac{\mu_0 \gamma_P \gamma_Q \hbar}{8\pi^3 r^3} \sum_{ij} A_{ij} \cdot \cos \varphi_i \cos \varphi_j \quad (3)$$

where A is the alignment tensor, r is the bond vector module, \hbar is the Planck constant, m_0 is the dielectric permittivity, and γ is the gyromagnetic radius. This formula requires the determination of the orientation of the protein in the alignment medium, which was done using a structure-based method (see Structure-Based Calculation of the Alignment Tensors Section). Although we have used a reference ensemble approach, and thus in principle we are able to define reference RDCs for all bond vectors in the protein, we have used only NH and CN bond vectors, which are the most commonly measured RDCs. To ensure a realistic case, we randomly removed 17% of the RDCs from the loop or terminal regions of the protein, including glycine residues, which usually give rise to signal broadening, and proline residues, which do not have an NH group. In total we employed 610 RDC values. In addition, a random perturbation of 0.15 Hz was applied on the final calculated values of the reference RDCs in order to account for experimental errors.

RDCs Calculation in the Restrained Simulations. The restrained simulations were carried out by using multiple replicas, a procedure that requires an averaging of the RDCs over the M replicas:

$$D_{\text{PQ}}^{\text{Calc}} = -\frac{\mu_0 \gamma_P \gamma_Q \hbar}{8\pi^3 r^3} \cdot \frac{1}{M} \sum_l \sum_{ij} A_{ij,l} \cdot \cos \varphi_{i,l} \cos \varphi_{j,l} \quad (4)$$

where l runs over the M replicas.

■ ASSOCIATED CONTENT

S Supporting Information. Figures S1–S6. This material is available free of charge via the Internet at <http://pubs.acs.org>.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: mv245@cam.ac.uk

■ ACKNOWLEDGMENT

This work was supported by the EPSRC (A.D.S.), the BBSRC (R.W.M. and M.V.), and the Wellcome Trust (M.V.).

■ REFERENCES

- (1) Russel, D.; Lasker, K.; Phillips, J.; Schneidman-Duhovny, D.; Velazquez-Muriel, J. A.; Sali, A. *Curr. Opin. Cell Biol.* **2009**, *21*, 97–108.
- (2) Henzler-Wildman, K. A.; Thai, V.; Lei, M.; Ott, M.; Wolf-Watz, M.; Fenn, T.; Pozharski, E.; Wilson, M. A.; Petsko, G. A.; Karplus, M.; Hubner, C. G.; Kern, D. *Nature* **2007**, *450*, 838–844.
- (3) Boehr, D. D.; Nussinov, R.; Wright, P. E. *Nat. Chem. Biol.* **2009**, *5*, 789–796.
- (4) Mittermaier, A. K.; Kay, L. E. *Trends Biochem. Sci.* **2009**, *34*, 601–611.
- (5) Tsai, C. J.; Kumar, S.; Ma, B. Y.; Nussinov, R. *Protein Sci.* **1999**, *8*, 1181–1190.
- (6) Csermely, P.; Palotai, R.; Nussinov, R. *Trends Biochem. Sci.* **2010**, *35*, 539–546.
- (7) Eisenmesser, E. Z.; Millet, O.; Labeikovsky, W.; Korzhnev, D. M.; Wolf-Watz, M.; Bosco, D. A.; Skalicky, J. J.; Kay, L. E.; Kern, D. *Nature* **2005**, *438*, 117–121.
- (8) Mittermaier, A.; Kay, L. E. *Science* **2006**, *312*, 224–228.
- (9) Boehr, D. D.; Dyson, H. J.; Wright, P. E. *Chem. Rev.* **2006**, *106*, 3055–3079.
- (10) Salmon, L.; Bouvignies, G.; Markwick, P.; Blackledge, M. *Biochemistry* **2011**, *50*, 2735–2747.
- (11) Kalodimos, C. G. *Protein Sci.* **2011**, *20*, 773–782.
- (12) Tjandra, N.; Bax, A. *Science* **1997**, *278*, 1111–1114.
- (13) Tolman, J. R.; Flanagan, J. M.; Kennedy, M. A.; Prestegard, J. H. *Nat. Struct. Biol.* **1997**, *4*, 292–297.
- (14) Prestegard, J. H.; Al-Hashimi, H. M.; Tolman, J. R. *Q. Rev. Biophys.* **2000**, *33*, 371–424.
- (15) Meiler, J.; Prompers, J. J.; Peti, W.; Griesinger, C.; Bruschweiler, R. *J. Am. Chem. Soc.* **2001**, *123*, 6098–6107.
- (16) Bouvignies, G.; Markwick, P.; Bruschweiler, R.; Blackledge, M. *J. Am. Chem. Soc.* **2006**, *128*, 15100–15101.
- (17) Salmon, L.; Bouvignies, G.; Markwick, P.; Lakomek, N.; Showalter, S.; Li, D. W.; Walter, K.; Griesinger, C.; Bruschweiler, R.; Blackledge, M. *Angew. Chem. Intl. Ed.* **2009**, *48*, 4154–4157.
- (18) Hess, B.; Scheek, R. M. *J. Magn. Reson.* **2003**, *164*, 19–27.
- (19) Clore, G. M.; Schwieters, C. D. *J. Am. Chem. Soc.* **2004**, *126*, 2923–2938.
- (20) Clore, G. M.; Schwieters, C. D. *Biochemistry* **2004**, *43*, 10678–10691.
- (21) Lange, O. F.; Lakomek, N. A.; Fares, C.; Schroder, G. F.; Walter, K. F. A.; Becker, S.; Meiler, J.; Grubmuller, H.; Griesinger, C.; de Groot, B. L. *Science* **2008**, *320*, 1471–1475.
- (22) De Simone, A.; Richter, B.; Salvatella, X.; Vendruscolo, M. *J. Am. Chem. Soc.* **2009**, *131*, 3810–3811.
- (23) Huang, J. R.; Grzesiek, S. *J. Am. Chem. Soc.* **2010**, *132*, 694–705.
- (24) Esteban-Martin, S.; Fenwick, R. B.; Salvatella, X. *J. Am. Chem. Soc.* **2010**, *132*, 4626–4632.
- (25) Azurmendi, H. F.; Bush, C. A. *J. Am. Chem. Soc.* **2002**, *124*, 2426–2427.

- (26) Iwahara, J.; Zweckstetter, M.; Clore, G. M. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 15062–15067.
- (27) Showalter, S. A.; Bruschiweiler, R. *J. Am. Chem. Soc.* **2007**, *129*, 4158–4159.
- (28) Marsh, J. A.; Forman-Kay, J. D. *J. Mol. Biol.* **2009**, *391*, 359–374.
- (29) Stelzer, A. C.; Frank, A. T.; Bailor, M. H.; Andricioaei, I.; Al-Hashimi, H. M. *Methods* **2009**, *49*, 167–173.
- (30) Zweckstetter, M.; Bax, A. *J. Am. Chem. Soc.* **2000**, *122*, 3791–3792.
- (31) Almond, A.; Axelsen, J. B. *J. Am. Chem. Soc.* **2002**, *124*, 9986–9987.
- (32) Ferrarini, A. *J. Phys. Chem. B* **2003**, *107*, 7923–7931.
- (33) Wu, B.; Petersen, M.; Girard, F.; Tessari, M.; Wijmenga, S. S. *J. Biomol. NMR* **2006**, *35*, 103–115.
- (34) Zweckstetter, M. *Nat. Protoc.* **2008**, *3*, 679–690.
- (35) Louhivuori, M.; Otten, R.; Lindorff-Larsen, K.; Annala, A. *J. Am. Chem. Soc.* **2006**, *128*, 4371–4376.
- (36) Salvatella, X.; Richter, B.; Vendruscolo, M. *J. Biomol. NMR* **2008**, *40*, 71–81.
- (37) Kuriyan, J.; Petsko, G. A.; Levy, R. M.; Karplus, M. *J. Mol. Biol.* **1986**, *190*, 227–254.
- (38) Richter, B.; Gsponer, J.; Varnai, P.; Salvatella, X.; Vendruscolo, M. *J. Biomol. NMR* **2007**, *37*, 117–135.
- (39) Rasmussen, B. F.; Stock, A. M.; Ringe, D.; Petsko, G. A. *Nature* **1992**, *357*, 423–424.
- (40) Vitagliano, L.; Merlino, A.; Zagari, A.; Mazzarella, L. *Proteins* **2002**, *46*, 97–104.
- (41) Watt, E. D.; Shimada, H.; Kovrigina, E. L.; Loria, J. P. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 11981–11986.
- (42) Beach, H.; Cole, R.; Gill, M. L.; Loria, J. P. *J. Am. Chem. Soc.* **2005**, *127*, 9167–9176.
- (43) Boehr, D. D.; McElheny, D.; Dyson, H. J.; Wright, P. E. *Science* **2006**, *313*, 1638–1642.
- (44) Lindorff-Larsen, K.; Best, R. B.; DePristo, M. A.; Dobson, C. M.; Vendruscolo, M. *Nature* **2005**, *433*, 128–132.
- (45) Vendruscolo, M. *Curr. Opin. Struct. Biol.* **2007**, *17*, 15–20.
- (46) Dobson, C. M. *Nature* **2003**, *426*, 884–890.
- (47) Palmer, A. G. *Chem. Rev.* **2004**, *104*, 3623–3640.
- (48) Markwick, P. R. L.; Bouvignies, G.; Salmon, L.; McCammon, J. A.; Nilges, M.; Blackledge, M. *J. Am. Chem. Soc.* **2009**, *131*, 16968–16975.
- (49) Korzhnev, D. M.; Religa, T. L.; Banachewicz, W.; Fersht, A. R.; Kay, L. E. *Science* **2010**, *329*, 1312–1316.
- (50) Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. *J. Chem. Theory Comput.* **2008**, *4*, 435–447.
- (51) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. *Proteins* **2006**, *65*, 712–725.
- (52) Sorin, E. J.; Pande, V. S. *J. Comput. Chem.* **2005**, *26*, 682–690.
- (53) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926–935.
- (54) Scott, W. R. P.; Hunenberger, P. H.; Tironi, I. G.; Mark, A. E.; Billeter, S. R.; Fennel, J.; Torda, A. E.; Huber, T.; Kruger, P.; van Gunsteren, W. F. *J. Phys. Chem. A* **1999**, *103*, 3596–3607.
- (55) Berendsen, H. J. C.; Grigera, J. R.; Straatsma, T. P. *J. Phys. Chem.* **1987**, *91*, 6269–6271.
- (56) Wlodawer, A.; Svensson, L. A.; Sjolín, L.; Gilliland, G. L. *Biochemistry* **1988**, *27*, 2705–2717.
- (57) Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. J. *Comput. Chem.* **1997**, *18*, 1463–1472.
- (58) Darden, T.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993**, *98*, 10089–10092.
- (59) Berisio, R.; Lamzin, V. S.; Sica, F.; Wilson, K. S.; Zagari, A.; Mazzarella, L. *J. Mol. Biol.* **1999**, *292*, 845–854.
- (60) Bax, A. *Protein Sci.* **2003**, *12*, 1–16.

Calculation of the Absolute Free Energy of Binding and Related Entropies with the HSMD-TI Method: The FKBP12-L8 Complex

Ignacio J. General, Ralitsa Dragomirova, and Hagai Meirovitch*

Department of Computational and Systems Biology, University of Pittsburgh School of Medicine, 3059 BST3, Pittsburgh, Pennsylvania 15260, United States

ABSTRACT: The hypothetical scanning molecular dynamics (HSMD) method is used here for calculating the absolute free energy of binding, ΔA^0 , of the complex of the protein FKBP12 with the ligand SB2 (also denoted L8)—a system that has been studied previously for comparing the performance of different methods. Our preliminary study suggests that considering long-range electrostatics is imperative even for a hydrophobic ligand such as L8. Therefore, the system is modeled by the AMBER force field using Particle Mesh Ewald (PME). HSMD consists of three stages applied to both the ligand–solvent and ligand–protein systems. (1) A small set of system configurations (frames) is extracted from an MD trajectory. (2) The entropy of the ligand in each frame is calculated by a reconstruction procedure. (3) The contribution of water and protein to ΔA^0 is calculated for each frame by gradually increasing the ligand–environment interactions from zero to their full value using thermodynamic integration (TI). Unlike the conventional methods, the structure of the ligand is kept fixed during TI, and HSMD is thus free from the end-point problem encountered with the double annihilation method (DAM). Therefore, the need for applying restraints is avoided. Furthermore, unlike the conventional methods, the entropy of the ligand and water is obtained directly as a byproduct of the simulation. In this paper, in addition to the difference in the internal entropies of the ligand in the two environments, we calculate for the first time the *external* entropy of the ligand, which provides a measure for the size of the active site. We obtain $\Delta A^0 = -10.7 \pm 1.0$ as compared to the experimental values -10.9 and -10.6 kcal/mol. However, a protein/water system treated by periodic boundary conditions grows significantly with increasing protein size, and the computation of ΔA^0 would become expensive by all methods. Therefore, we also apply HSMD to FKBP12-L8 described by the GSBP/SSBP model of Roux's group (implemented in the software CHARMM) where only part of the protein and water around the active site are considered and long-range electrostatic effects are taken into account. For comparison, this model was also treated by the double decoupling method (DDM). The two methods have led to comparable results for ΔA^0 , which are somewhat lower than the experimental value. The ligand was found to be more confined in the active site described by GSBP/SSBP than by PME, where its entropy in solvent is larger than in the active site by 1.7 and by 5.5 kcal/mol, respectively.

I. INTRODUCTION

The central aim of this paper is to further develop our hypothetical scanning molecular dynamics (HSMD) method for calculating the absolute free energy of binding. To emphasize the significance of our method, we first discuss the importance of the problem and the properties of existing techniques.

The development of simulation methods for calculating the affinity (free energy) of molecular binding is important for both academic and practical reasons as tools for elucidating the mechanisms of complex biological processes, such as the action of hormones, the recognition of antigens by the immune system, the catalysis of chemical reactions by enzymes, and the action of drugs. Therefore, such methods can be used in rational drug design, leading to various therapeutic means. It is important in particular to devise highly accurate methods for calculating the free energy of binding based on detailed molecular interactions and rigorous statistical mechanics; such methods are required in the refinement stage of screening procedures based on simplified (fast) scoring functions. Of special interest in this category are methods for calculating the *absolute* (standard) free energy of binding of a ligand to a protein.

A great deal of work has been done in this direction, where various techniques have been developed and applied to a wide

range of problems (see refs 1–22 and references cited therein). A central *rigorous* approach is based on thermodynamic cycles, where the interactions between the ligand and its environment are decreased to zero in both, the active site and the bulk solution (water), using thermodynamic integration (TI) or free energy perturbation (FEP) procedures. However, this approach, called the double annihilation method (DAM),^{1–5} is not straightforward since during the final stages of TI the ligand leaves the active site and starts wandering within the volume, which makes it extremely difficult to obtain converged results; also in most of the DAM applications, the effect of the standard state was not applied (see ref 12 as a recent example). Still, this method has been used successfully in recent years,^{8,12} where in a later study¹³ the effect of the standard state has been considered.

The end-point problem has been rigorously solved by adding restraints which hold the ligand in the active site; the corresponding bias introduced is later removed by releasing the restraints. Because of the additional integration step involved, this procedure is sometimes called the double decoupling method (DDM).^{3–5} DDM has been developed systematically in the past

Received: July 13, 2011

Published: October 27, 2011

15 years,^{7,9,15} where various implementation issues have been improved and a large number of complexes have been successfully studied (see review in ref 9). Historically, a (translational) restraining potential was introduced rigorously first by Hermans and Shankar¹ and extended later to a more complex system by Roux's group.¹⁸ Hermans and Wang were also the first to introduce an angular restraint,¹⁹ and more restraints were added later by Roux and Deng^{20,9} (see also refs 4 and 5).

Therefore, it is desirable to enrich the arsenal of methods available in this field by developing *alternative* techniques that are free from the problems discussed above and provide, in addition to the binding free energy, other information, such as the entropy. A step in this direction has been made in our recent publication, where HSMD has been extended for the first time to a binding problem, treating the avidin–biotin complex.²² HSMD is an *exact* technique for calculating the *absolute* entropy and free energy, which has been developed in our group over the past several years and has been tested systematically on models of increasing complexity, liquid argon,²³ TIP3P water,²³ self-avoiding walks,²⁴ and peptides,²⁵ where HSMD results were compared to those obtained by well established techniques (more accurately, HSMC was used in these calculations where Monte Carlo replaces MD). While HSMD is a general technique, which is applicable to any system (including fluids), for a protein in explicit water, the contribution of water to the free energy is calculated much more efficiently by a TI (or FEP) procedure that has been incorporated within the framework of HSMD; the combined method—called HSMD-TI (or HSMD-FEP)—has been applied successfully to several mobile loops in proteins.^{26–30}

HSMD-TI consists of three stages applied to both the ligand–solvent and ligand–protein systems. (1) A small set of system configurations (frames) is extracted from an MD trajectory. (2) The ligand's entropy in each frame is calculated by a reconstruction procedure. (3) The contribution of water and protein to the binding free energy is calculated for each frame by gradually increasing the ligand–environment interactions from zero to their full value using TI. Because during TI the structure of the ligand is kept fixed, the end-point problem (of DAM) does not exist and the need for applying restraints (DDM) is avoided. Furthermore, unlike DAM and DDM, several partial entropies of the system are obtained directly as byproducts of the simulation, and the method is, in particular, suitable to handle large flexible ligands. In this study, we calculate for the first time the external entropy of the ligand in the active site, which constitutes a measure for the global movement of the bound ligand, thereby providing estimation for the size of the active site. All of these entropies provide microscopic insights into the binding mechanism and are thus important in rational drug design.

In our initial avidin–biotin study²² (using TINKER³¹), we were mainly interested in checking the performance of HSMD-TI; therefore, the modeling has been somewhat limited. In particular, only a spherical part of the tetrameric avidin, consisting of residues within a distance of 18 Å from the center of the binding site, were considered, and in most of the calculations, this part (the template) was kept fixed during the simulations. Also, a comparable water sphere was used. Finally, long-range electrostatic effects were ignored. Still, including the contribution of a mobile loop has led to the experimental absolute free energy of binding.

The aim of this paper is to develop HSMD-TI further by applying it to the protein FKBP12 bound to the ligand SB3 (see below), where this system is described by models of various complexities. FKBP12 is a small protein (107 residues), which

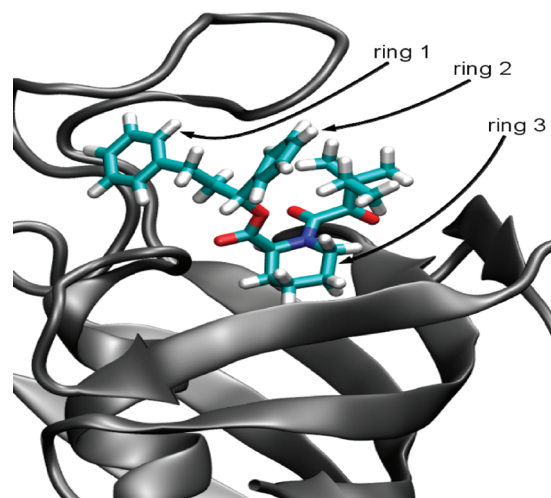


Figure 1. Illustration of the FKBP12–L8 complex. The three rings of the ligand are numbered; ring 3 is positioned at the bottom of the active site.

catalyzes the *cis*–*trans* isomerization of peptidyl–prolyl bonds.³² FK506 is a key drug used for immunosuppression in organ transplant. It binds strongly to FKBP12,³³ and the FKBP12/FK506 complex, in turn, binds and inhibits calcineurin, thus blocking the signal transduction pathway for the activation of T-cells.^{34,35}

Crystal structures of FKBP12 in complex with several ligands are available,^{36–38} and the binding constants of the FK506-related ligands with FKBP12 have been experimentally determined.³⁷ Therefore, this system has served as a rich platform to test and validate different computational strategies for estimating binding free energies.^{8,11,13,39–42} In particular, a set of eight ligands complexed with FKBP12 (out of the 27 studied in ref 37) have become the target for absolute free energy calculations by several groups who applied different methods. Wang et al.⁴⁰ have used DDM with their models—Spherical Solvent Boundary Potential (SSBP⁴³) and Generalized Solvent Boundary Potential (GSBP⁴⁴). GSBP is an example of what we call in this paper a “model of partial structure”, i.e., a model where only part of the protein close to the active site is considered, and this part is covered by a relatively small sphere of water. Notice, however, that both GSBP and SSBP take into account long-range electrostatics (reaction field). Pande's group and Fujitani et al. used DAM,^{8,11,13,39} where the (entire) complex and the ligand in the bulk were immersed in a box of explicit water and long-range electrostatic effects were taken into account by periodic boundary conditions with Particle Mesh Ewald (PME).⁴⁵

We have decided to study the SB3 ligand, which is also called L8 (68 atoms, Figure 1), because in the set of eight ligands, it is the smallest for which the crystal structure of the complex is known. As discussed later in section III.1, in preliminary studies, L8–FKBP12 was described by a model of partial structure handled by the AMBER 10 package⁴⁶ with an empirical potential energy—the AMBER force field,⁴⁷ TIP3P water,⁴⁸ and the General AMBER Force Field (GAFF)⁴⁹ for the ligand. Because the ligand is hydrophobic, we had reasons to believe that applying long-range electrostatic interactions would not be necessary (see section III.1); however, the results were found to be unsatisfactory. Therefore, we have decided to treat the FKBP–L8 system (from now on we omit “12”) with two models that consider long-range electrostatics—PME implemented in the software package

AMBER⁴⁶ and the SSBP/GSBP^{43,44} modeling, which is incorporated within the program CHARMM.⁵⁰

II. THEORY AND METHODOLOGY

II.1. Theory of Binding. Imagine a *dilute* solution of a protein (P) and a ligand (L) in a volume V in equilibrium with their complex (PL), $P + L \leftrightarrow PL$. The equilibrium constant, K_b , is defined by the equilibrium concentrations (denoted $[]$) of these components, $K_b = [LP]/[P][L]$, where K_b leads to the absolute (standard) free energy of binding, ΔA^0 . Since our system is defined in the NVT ensemble, we obtain^{4,5}

$$\Delta A^0 = -k_B T \ln \frac{V Z_{PL,N} Z_{0,N}}{V^0 Z_{P,N} Z_{L,N}} = -k_B T \ln \frac{\bar{Z}_{PL,N} Z_{0,N}}{8\pi^2 V^0 \bar{Z}_{P,N} \bar{Z}_{L,N}} \quad (1)$$

where T is the absolute temperature, k_B is the Boltzmann constant, and N stands for the number of solvent molecules (water). $Z = V 8\pi^2 \bar{Z}$, where $\bar{Z}_{PL,N}$, $\bar{Z}_{P,N}$, and $\bar{Z}_{L,N}$ are the conformational partition functions of the complex, protein, and ligand all in water; $Z_{0,N}$ is the partition function of N water molecules in the volume. The bar means that P and L in \bar{Z} are defined by *internal coordinates*, where the integration ($V 8\pi^2$) over the external coordinates (e.g., a reference atom and three Euler angles defined by two more atoms) has already been carried out, where V is the system's volume and $V^0 = 1660 \text{ \AA}^3$ is the standard volume. ΔA^0 is expressed in terms of \bar{Z} because \bar{Z} does not depend explicitly on V , and with HSMD-TI we mainly calculate internal entropies and free energies. Notice that the ligand moves in the active site; i.e., $\bar{Z}_{PL,N}$ includes a localized ligand partition function where its coordinates can also be divided into internal and external, where the contribution of the latter will be calculated by HSMD (rather than analytically as in the solvent). ΔA^0 is expressed in terms of configurational (Helmholtz) free energies, $F = -k_B T \ln \bar{Z}$ (and $F_{0,N} = -k_B T \ln Z_{0,N}$) and an additional term

$$\begin{aligned} \Delta A^0 &= (F_{PL,N} - F_{P,N}) - (F_{L,N} - F_{0,N}) + k_B T \ln(8\pi^2 V^0) \\ &= \Delta F_p - \Delta F_{\text{sol}} + k_B T \ln(8\pi^2 V^0) \end{aligned} \quad (2)$$

ΔF_p and ΔF_{sol} are free energy differences defined for the protein and solvent environments, respectively, which are calculated by HSMD-TI. Also, the absolute Gibbs free energy $\Delta G^0 \sim \Delta A^0$, since $\Delta G^0 = \Delta A^0 + P^0 \Delta \bar{V}_{PL}$ where $P^0 \Delta \bar{V}_{PL}$ is small and can be neglected.^{4,5}

II.2. The Philosophy of HSMD. The HSMD method (as well as HSMC) enables one to calculate the *absolute* entropy and free energy from a sample generated by MD,^{51,52} MC,⁵³ or any other simulation technique. HSMD is based on the fact that a system configuration can, in principle, be generated *exactly* also by a step-by-step (growth) procedure, where particles are added gradually to an initially empty volume using transition probabilities (TPs). A trivial example is an *ideal chain* of N bonds on a square lattice, i.e., a chain without the excluded volume interaction. An ideal chain can be simulated *exactly* by the dynamical MC method where the entropy, however, is unknown; alternatively, a chain configuration, i , can be constructed (from nothing) as a random walk where a bond's direction (out of four) is chosen with $TP = 1/4$ and added at each step. Here, the Boltzmann probability is known, $P_i^B = (1/4)^N$, and thus the entropy is

known as well ($S = -k_B \ln P_i^B$). Clearly, large samples constructed by MC or as random walks are equivalent in the sense that they lead to the same thermodynamic averages and fluctuations.

While an exact growth procedure (called the exact scanning method^{54,30}) can be defined for any system (e.g., water, protein in water), application of such a procedure will in general be very inefficient. However, relying on the equivalence among exact simulation methods mentioned above, one can assume that a given MD (MC) sample has rather been generated by the exact scanning method, which enables one to reconstruct each conformation i by calculating the TP densities that *hypothetically* were used to create it step-by-step. Application of this *hypothetical scanning* (HS) procedure is much more efficient than the (direct) scanning method.

In practice, the product of the TPs leads to an approximation, P_i , for the correct Boltzmann probability P_i^B , where from P_i various free energy functionals (F) can be defined. However, no inherent approximation is applied in the calculation of the TPs. That is, all of the system interactions are taken into account, and the only approximation involved is due to insufficient MD sampling for their calculation. In this respect, HSMD is considered to be an exact method.²³ More specifically, we calculate $S(n_f) = -k_B \sum P_i^B \ln P_i(n_f)$, which constitutes a rigorous upper bound for the correct entropy S ; the larger the sample size, n_f , the better is P_i and the smaller is $S(n_f)$. Thus, one can follow the convergence of $S(n_f)$ to S (from above) as n_f is increased until the increase in $S(n_f)$ becomes smaller than the statistical error; i.e., $S(n_f)$ is *exact* within this error (see detailed discussions related to eqs 4, 6, and 7 below). In practice, this convergence will not be attained satisfactorily (even though in a recent publication, results for $S(n_f)$ of biotin were found to be close to convergence). However, we are not interested in $S(n_f)$ itself but in the entropy difference $[\Delta S(n_f)]$, where in the application for binding, it is between the ligand's entropy in solvent and in the active site. Typically, $\Delta S(n_f)$ converges very rapidly as a function of n_f (due to cancellation of errors) with a statistical error that is smaller than other errors involved in the calculation of ΔA^0 (see refs 22–30 and a note in ref 55). HSMD (as well as HSMC) has unique features: it provides rigorous lower and upper bounds for F , which enables one to determine the accuracy from HSMD results alone without the need to know the correct answer (however, not all of these features are used in the present application). Furthermore, F can be obtained from a very small sample and in principle even from *any* single conformation (see next section, and ref 23).

II.3. Fluctuations. The fact that HSMD provides an approximation for P_i^B means that the *absolute* entropy can be directly obtained and that the free energy, F , can in principle be calculated from a single conformation and in practice from a small sample. This stems from the fact that if the energy, E_i , and $P_i^B (= \exp(-E_i/k_B T)/Z)$ of *any* structure i are known, the *total* free energy of the system, F , is known as well, since $F_i = E_i + k_B T \ln P_i^B = E_i + k_B T(-E_i/k_B T - \ln Z) = -k_B T \ln Z = F$; in other words, the *exact* F has zero fluctuation.^{56,57} (While this may look strange, one should bear in mind that the calculation of P_i^B for a single conformation depends on the entire ensemble through its normalization factor, which is the partition function.) Still, for an approximate P_i the fluctuation of $F(P_i)$ is finite, but it is expected to decrease as the approximation improves, i.e., as $P_i \rightarrow P_i^B$ where the required sample size decreases as well.^{56,57,23} Notice, however, that unlike the free energy discussed above, the entropy (and energy) cannot

be obtained from an *arbitrary* single structure but should be calculated from a Boltzmann sample. We shall return to this matter in section II.6.

II.4. The Reconstruction Procedure for Binding. The process starts by carrying out two *production* MD runs of the ligand in water and the protein–ligand complex in water. From these trajectories, we determine two sets of n_s equally spaced frames (snapshots) for later reconstruction and TI calculations by HSMD-TI. As discussed earlier, in the reconstruction process, one seeks to calculate the Boltzmann probability density P_i^B related to each of the frames, as a product of transition probabilities (TPs). For simplicity, we describe this procedure as applied to the ligand in water, where the ligand can be any organic molecule (e.g., L8). For this system (ligand in water), it is convenient to reconstruct the ligand conformation first, followed by the reconstruction of the configuration, \mathbf{x}^N , of the N water molecules.^{22,26–30}

In the next step, we identify the K (internal) dihedral and bond angles ordered along the chain where their set is denoted by $[\alpha_k]$, $k = 1 \dots K$. We then denote the ordered heavy atoms (and polar hydrogens) along the chain by $k' = 1 \dots K/2$; we shall see below that in the reconstruction process the position of atom k' is determined solely by the dihedral and bond angles, $k - 1$ and k ($k = 2k'$), respectively, and the bond length r_k with the Jacobian $r_k^2 \sin \alpha_k$. The calculation of the TPs depends on the fluctuations in these three coordinates. However, bond stretching is ignored because in general it contributes very little to entropy differences^{58–60} (notice, however, that bond stretching can straightforwardly be implemented within HSMD; see eq 9 below); therefore, in practice, the only variables considered are α_{k-1} and α_k , and for the sake of completeness in the calculations, we use the Jacobian $\langle r \rangle^2 \sin \alpha_k$, where the constant $\langle r \rangle = 1.6 \text{ \AA}$ is an average value that appears for the ligand in the protein and the solvent environments and thus gets canceled in entropy differences. We then calculate the variability range

$$\Delta\alpha_k = \alpha_k(\max) - \alpha_k(\min) \quad (3)$$

where $\alpha_k(\max)$ and $\alpha_k(\min)$ are the maximum and minimum values of α_k found in each sample.

Each of the system configurations (frames; $[\alpha_k], \mathbf{x}^N$; denoted i for brevity) is reconstructed in two stages, where the ligand structure is reconstructed first followed by the reconstruction of the water configuration. Because the position of atom k' is defined by a dihedral and a bond angle, one has to calculate their TP simultaneously. Thus, at step k' ($k = 2k'$) of stage 1, $k - 2$ angles $\alpha_{k-2} \dots \alpha_1$ have already been reconstructed, and the TP density of $\alpha_{k-1} \alpha_k$, $\rho(\alpha_{k-1} \alpha_k | \alpha_{k-2}, \dots, \alpha_1)$, is calculated from an MD run, where the *entire future* of the ligand and water is moved [i.e., ligand's atoms $k', k' + 1, \dots, K/2$ and their connected hydrogens, and all the water molecules] while the past (ligand's atoms 1, 2, ..., $k' - 1$ and their connected hydrogens) are held fixed at their values in conformation i (see Figure 2 in ref 29 and a note in ref 61). By considering a future conformation every 20 fs, a sample of size n_f is generated. Two small segments (bins) $\delta\alpha_{k-1}$ and $\delta\alpha_k$ are centered at $\alpha_{k-1}(i)$ and $\alpha_k(i)$, respectively, and the number of *simultaneous* visits, n_{visit} of the future chain to these two bins during the simulation is calculated. One obtains^{22,26–30}

$$\begin{aligned} \rho_{\text{ligand}}(\alpha_{k-1} \alpha_k | \alpha_{k-2}, \dots, \alpha_1) &\approx \rho^{\text{HS}}(\alpha_{k-1} \alpha_k | \alpha_{k-2}, \dots, \alpha_1) \\ &= n_{\text{visit}} / [n_f \delta\alpha_{k-1} \delta\alpha_k J] \end{aligned} \quad (4)$$

where $\rho^{\text{HS}}(\alpha_{k-1} \alpha_k | \alpha_{k-2}, \dots, \alpha_1)$ becomes exact for very large n_f ($n_f \rightarrow \infty$) and very small bins ($\delta\alpha_{k-1}, \delta\alpha_k \rightarrow 0$). This means that in practice $\rho^{\text{HS}}(\alpha_{k-1} \alpha_k | \alpha_{k-2}, \dots, \alpha_1)$ will be approximate due to insufficient future sampling (finite n_f) and relatively large bins (where their optimal size depends on n_f). The Jacobian is $J = \langle r \rangle^2 \sin \alpha_k$. The corresponding probability density related to the ligand's conformation is

$$\rho^{\text{HS}}(\alpha_K, \dots, \alpha_1) = \rho^{\text{HS}}([\alpha_k]) = \prod_{k=1,2}^{K-1} \rho^{\text{HS}}(\alpha_k \alpha_{k+1} | \alpha_{k-1}, \dots, \alpha_1) \quad (5)$$

$\rho^{\text{HS}}([\alpha_k])$ defines an approximate entropy functional denoted S_{ligand}^A , which can be shown (using Jensen's inequality, see ref 23) to constitute a *rigorous* upper bound for the correct S_{ligand}

$$S_{\text{ligand}}^A = -k_B \int_m \rho_{\text{ligand}}^B([\alpha_k]) \ln \rho^{\text{HS}}([\alpha_k]) J d[\alpha_k] \quad (6)$$

$\rho_{\text{ligand}}^B([\alpha_k])$ is the Boltzmann probability density of $[\alpha_k]$, and J is the Cartesian to the internal coordinates Jacobian. ($S_{\text{ligand}}^A \geq S_{\text{ligand}}$ is also known as the Gibbs' inequality). Being an upper bound suggests that S_{ligand}^A will decrease as the approximation improves. It should be noted that S_{ligand}^A is defined by our procedure and can be viewed as the conformational internal “entropy of mean force”, which constitutes a measure of a pure geometrical character for the flexibility of the ligand; it is estimated by $\bar{S}_{\text{ligand}}^A$ from an MD (Boltzmann) sample of size n_s

$$\bar{S}_{\text{ligand}}^A = -(k_B/n_s) \sum_{t=1}^{n_s} \ln \rho^{\text{HS}}(t) \quad (7)$$

Notice again that S_{ligand}^A is the internal entropy of the ligand. The internal entropy of the ligand in the active site is calculated in the same way, where, however, the future includes the waters as well as the protein atoms, which are all moved by MD in the reconstruction process. We denote the entropies of the ligand in the protein and the solvent (water) environments by $S_{\text{ligand}}^A(\text{p})$ and $S_{\text{ligand}}^A(\text{sol})$, respectively, where our main interest is in their *converged* difference ΔS_{ligand} , which is expected to be the *exact* difference within the statistical error (see discussion in section II.2 and ref 55)

$$\Delta S_{\text{ligand}} = S_{\text{ligand}}^A(\text{sol}) - S_{\text{ligand}}^A(\text{p}) \text{ converged} \quad (8)$$

Thus, we calculate $S_{\text{ligand}}^A(\text{sol})$ and $S_{\text{ligand}}^A(\text{p})$ for increasing n_f and decreasing bins, verifying that both entropies decrease monotonically as the approximation improves; i.e., both approach the correct values from above (notice again that the commanding parameter is n_f where the bin size should correspond to the given statistics (n_f), and it cannot be decreased independently). Typically, the convergence of $\Delta S_{\text{ligand}}^A$ is much faster than that of the individual entropies, due to cancellation of comparable errors in $S_{\text{ligand}}^A(\text{sol})$ and $S_{\text{ligand}}^A(\text{p})$. Thus, one can obtain ΔS_{ligand} in a desired accuracy, when the changes in the improved values of $\Delta S_{\text{ligand}}^A$ are smaller than a given statistical error (notice that this error also depends on the sample size, n_s , and other simulation parameters). The range of errors obtained in our previous work has been 0.2–1 kcal/mol. Therefore, HSMD is considered to be an exact procedure.

Two comments should be made. First, besides S_{ligand} , one should also include in ΔA^0 (eq 2), the contribution of the ligand–ligand energy, $E_{\text{ligand–ligand}}$, which is averaged over the samples of the two environments (of size n_s) leading to

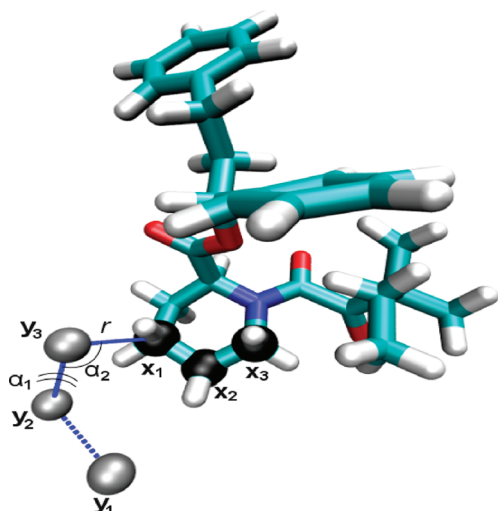


Figure 2. Reconstruction of the external coordinates. \mathbf{x}_1 , \mathbf{x}_2 , and \mathbf{x}_3 are the coordinates of three (successive) reference atoms on ring 3 of the ligand; \mathbf{y}_1 , \mathbf{y}_2 , and \mathbf{y}_3 are three fixed positions in space. The position \mathbf{x}_1 is defined by “dihedral angle” α_1 , “bond angle” α_2 , and $r = |\mathbf{x}_1 - \mathbf{y}_3|$. In the reconstruction of the internal coordinates the right branch is treated first.

$E_{\text{ligand-ligand}}(\text{sol})$ and $E_{\text{ligand-ligand}}(\text{p})$. Second, notice that the effect of bond stretching can be considered by defining a bin δr_k around the existing k 'th bond length, r_k (with the related Jacobian, $J = rk^2 \sin \alpha_k$), and eq 4 becomes

$$\rho_{\text{ligand}} = n_{\text{visit}} / [n_f \delta \alpha_{k-1} \delta \alpha_k \delta r_k J] \quad (9)$$

where n_{visit} is the number of *simultaneous* visits of the future chain to the *three* bins.

II.5. Entropy of the External Coordinates. To calculate the external entropy within the framework of HSMD, one defines initially three successive atoms of the ligand (with coordinates \mathbf{x}_1 , \mathbf{x}_2 , and \mathbf{x}_3 in the laboratory frame) where the TPs for conformation i are based on a small cube (bin) of volume V_{cube} around \mathbf{x}_1 and a small bin for each of the Euler angles. More specifically, to be consistent with the reconstruction of the internal coordinates, the position of atom 1 (\mathbf{x}_1) is expressed by spherical coordinates. Thus, one determines three fixed points in space with coordinates \mathbf{y}_1 , \mathbf{y}_2 , and \mathbf{y}_3 , which together with \mathbf{x}_1 define a “dihedral angle” α_1 (based on \mathbf{y}_1 , \mathbf{y}_2 , \mathbf{y}_3 , and \mathbf{x}_1), a “bond angle” α_2 (based on \mathbf{y}_2 , \mathbf{y}_3 , and \mathbf{x}_1), a distance $r = |\mathbf{x}_1 - \mathbf{y}_3|$, and their corresponding bins (see Figure 2); these bins together with the Jacobian, $r^2 \sin(\alpha_2)$, define $V_{\text{cube}}(i)$ of ligand conformation i . The contribution of atom 1 to S_{external} [denoted $S_{\text{external}}(1)$] is based on eq 9, where the occupancy, n_{visit} of \mathbf{x}_1 in $V_{\text{cube}}(i)$ is obtained by generating n_f future conformations of the entire ligand (and environment); $S_{\text{external}}(1)$ is averaged over the sample of n_s frames. Notice that for the ligand in solvent such calculation would be very time-consuming and not necessary because the result $\text{TP} = 1/V$ is known analytically.

After atom 1 has been reconstructed, it becomes fixed at its position in conformation i and the contribution of atom 2 is calculated. Thus (assuming that $r = |\mathbf{x}_2 - \mathbf{x}_1|$ is constant), one defines a dihedral angle α_3 (based on \mathbf{y}_2 , \mathbf{y}_3 , \mathbf{x}_1 , and \mathbf{x}_2) and a bond angle α_4 (based on \mathbf{y}_3 , \mathbf{x}_1 , and \mathbf{x}_2) with their bins and the Jacobian, $r^2 \sin(\alpha_4)$, where $r^2 = 1$. [$r^2 = 1$ because we calculate only the angular contribution, i.e., for a partial spherical surface of radius $r = 1$; this is compatible with the solvent side, where a (very

long) such simulation would lead to the (analytically known) $\text{TP} = 1/4\pi$.] On the protein side, one obtains $S_{\text{external}}(2)$ through eq 4.

Finally, the contribution of atom 3 is considered by defining the pair α_5 (\mathbf{y}_3 , \mathbf{x}_1 , \mathbf{x}_2 , and \mathbf{x}_3) and α_6 (\mathbf{x}_1 , \mathbf{x}_2 , and \mathbf{x}_3) (assuming that $r = |\mathbf{x}_3 - \mathbf{x}_2|$ is constant). However, only the effect of the dihedral α_5 is considered (using eq 4) by defining an artificially large bin for the bond angle α_6 where the corresponding contribution is $S_{\text{external}}(3)$. Here, the Jacobian is $r^2 \sin(\alpha_6) = 1$, which is again compatible with the solvent side where a long reconstruction simulation would lead to the analytically known $\text{TP} = 1/2\pi$.

One can define the external entropy for the ligand in solvent as $-k_B T \ln[1/(V8\pi^2)]$ where $1/(V8\pi^2)$ is the product of the above three TPs; however, this entropy is already included in eq 1. Correspondingly, the sum of the three external entropies calculated above for the ligand in the protein environment leads to S_{external}

$$S_{\text{external}} = S_{\text{external}}(1) + S_{\text{external}}(2) + S_{\text{external}}(3) \quad (10)$$

where S_{external} constitutes a measure for the conformational freedom of the ligand in the active site; hence, it provides some estimation for the size of the latter.

It should be pointed out that because with HSMD the whole future is scanned, the reconstruction procedure can in principle be started from any atom, i.e., from the “first” atom of a chain-like ligand, from the “last” one (in an opposite direction), or from any middle atom, where the order in which the two branches are reconstructed is arbitrary. In practice, a large enough n_f is expected to lead to close results for $S_{\text{ligand}}^A + S_{\text{external}}$ based on different starting atoms (this choice of atoms might lead to different computational efficiencies; however, this issue has not been studied by us as yet). Also, because the reconstruction process treats all atoms in a successive manner, the three reference atoms should be nearest neighbors along the ligand, and the position of the first atom should preferably be chosen in such a way that S_{external} will best express the global movement of the ligand in the active site. While an ideal first (global) coordinate would be the center of mass, this coordinate is inadequate due to its constant change during reconstruction. We shall see later that to achieve this goal the reference atoms of L8 are defined close to the middle of the chain.

II.6. HSMD-TI. After a ligand conformation i (say, in solvent) has been reconstructed, one should reconstruct the water configuration $\mathbf{x}^N(i)$ in the presence of a *fixed* ligand structure,²³ which would lead to $\rho_{\text{water}}(i)$ and to the contribution, F_i , to the free energy, F :

$$F_i = [E_{\text{water-water}}(i) + E_{\text{water-ligand}}(i) - k_B T \ln \rho_{\text{water}}(i)] \\ + [E_{\text{ligand-ligand}}(i) - TS_{\text{ligand}}^A(i)] = F_{\text{water}}(i) + F_{\text{ligand}}(i) \quad (11)$$

where, for large n_f [defined also for $\rho_{\text{water}}(i)$], $F_i \rightarrow F$ (see discussion in section II.3). However, reconstructing $\rho_{\text{water}}(i)$ would be a time-consuming process that would be necessary for calculating the absolute value of $F_{L,N}$ (eq 2). Nevertheless, $F_{\text{water}}(i)$ can also be obtained by a TI procedure where an ideal gas (in N , V , and T) is gradually transformed first to liquid water, leading to $F_{0,N}$ (eq 2); then the ligand–water interactions are turned on and are integrated to their full value, leading to $F_i^{\text{TI}}(\text{sol})$, which for a long integration is equal to $F_{\text{water}}(i)$ (eq 11). Therefore,

the water contribution to the difference $\Delta F_{\text{sol}} = F_{L,N} - F_{0,N}$ (eq 2) can be calculated for each of the n_s configurations i by a TI procedure in which the already reconstructed ligand conformation $[\alpha_k]$ is kept fixed and the water–ligand interactions are increased gradually from zero to their full values, leading to $F_i^{\text{TI}}(\text{sol})$; in practice, however, it is easier to decrease these interactions to zero (using $\lambda \rightarrow 0$), obtaining the negative of $F_i^{\text{TI}}(\text{sol})$. This elimination of the ligand–environment interactions by TI is expected to be physically more reliable than building them from zero. In the latter case, one can encounter unphysical situations, where, for example, too many water molecules are trapped in the active site after integration. Thus, we carry out only the elimination TI but for quite a few different ligand structures, averaging the results where their standard deviation defines the error.

This suggests that, in principle, ΔF_{sol} can be obtained from any single configuration i by using large enough n_f for calculating $S_{\text{ligand}}(i)$ (eqs 6 and 7) and long enough integration, which leads to accurate $F_i^{\text{TI}}(\text{sol})$ (see discussion in section II.3). In practice, however, the results are averages over a relatively small sample size $10 \leq n_s \leq 100$, where the errors are defined by $\text{sd}/(n_s)^{1/2}$, where sd stands for the standard deviation (see section II.3); $\text{sd} \rightarrow 0$ as n_f and the integration time increase. In the protein environment, one obtains in the same way $F_i^{\text{TI}}(\text{p})$ by eliminating the ligand–protein and ligand–water interactions. Technically, the charges are eliminated first, followed by elimination of the Lennard-Jones (LJ) potential, using a soft-core potential. With TS_{external} , ΔA^0 (eq 2) becomes

$$\begin{aligned} \Delta A^0 &= \Delta F_{\text{p}} - \Delta F_{\text{sol}} + k_{\text{B}} T \ln(8\pi^2 V^0) - TS_{\text{external}} \\ &= [E_{\text{ligand-ligand}}(\text{p}) - TS_{\text{ligand}}(\text{p}) + F^{\text{TI}}(\text{p})] \\ &\quad - [E_{\text{ligand-ligand}}(\text{sol}) - TS_{\text{ligand}}(\text{sol}) + F^{\text{TI}}(\text{sol})] \\ &\quad + k_{\text{B}} T \ln(8\pi^2 V^0) - TS_{\text{external}} = \Delta E_{\text{ligand-ligand}} - T\Delta S_{\text{ligand}} \\ &\quad + \Delta F^{\text{TI}} + k_{\text{B}} T \ln(8\pi^2 V^0) - TS_{\text{external}} \quad (12) \end{aligned}$$

where all of the quantities (defined earlier) are averages over n_s snapshots and Δ denotes differences in the corresponding variables (protein–solvent); the errors of ΔA^0 and its different components are obtained from the standard deviations (sd) divided by $(n_s)^{1/2}$. Notice again that $V^0 8\pi^2$ and the term in the logarithm of S_{external} have the same dimensions, which disappear in the difference.

III. RESULTS AND DISCUSSION

III.1. Preliminary Studies without Long-Range Electrostatics. Because L8 is an uncharged hydrophobic ligand, we had reasons to believe that describing the complex with a good model of partial structure (where the whole protein is considered, it is covered with water, and cut-offs on nonbonded interactions are not imposed), long-range electrostatic interactions would not be significant, and they would get canceled in free energy differences. Previous binding studies of models of partial structure have been found to be successful.^{3,62}

Thus, we studied initially this system using the AMBER 10 package,⁴⁶ where due to its small size, the whole complex was considered (PDB code 1fkg, 1974 atoms). Around the active site, we defined a TIP3P⁴⁸ water sphere of 22 Å radius (658 molecules), which covered the whole protein (because the protein was not soaked in a large container of water, we consider this system a model of partial structure). For simulations of the solvated ligand, a sphere of 22 Å radius was used, containing 1310

waters. The protein was modeled by the AMBER99 force field⁴⁷ and the ligand by GAFF⁴⁹ (with the AM1-BCC partial charges). *No cut-offs* on nonbonded interactions were imposed. While this model includes a large part of the system, the HSMD-TI results obtained for the absolute free energy were too high (in the best case favoring unbinding by 5 kcal/mol) even after several attempts to optimize the structure of the ligand in the complex by minimizing the ligand–environment interaction energy prior to the production run. No improvement could be achieved also by charging the neutral histidines or adding chlorine ions to neutralize the total charge of the system.

As mentioned in the Introduction, these studies have led to the conclusion that long-range electrostatics should be considered, and below we describe the application of HSMD-TI to L8–FKBP modeled by periodic boundary conditions with PME. Results are also provided later for this system modeled by SSBP/GSBP.^{43,44}

III.2. Studies with Long-Range Electrostatics (PME). To apply PME, we also used the AMBER 10 package along with the AMBER99 and GAFF force fields (including TIP3P water) and the crystal structure described above. The unit cell of the periodic system was defined by constructing a truncated octahedron around the protein which was filled with 4581 water molecules. The minimum distance between the protein and the walls of the cell was 11 Å (this cell size is comparable to that used in previous studies of L8; also, to be compatible with these studies, the three histidine residues have positive charge). To neutralize the system, four chloride counterions were added at random locations. The periodic boundary conditions were defined by the PME algorithm,⁴⁵ with a cutoff of 10 Å. This system was optimized in several stages, starting with a short simulated annealing protocol consisting of energy minimization, where 10^4 steepest descent and 10^4 conjugate gradient steps were performed. Then, using MD, the system was heated to 600 K over 12 ps (using the Berendsen thermostat⁶³ with a 2 ps time constant), slowly cooled down to 100 K (28 ps), cooled down further (5 ps) with a smaller time constant (1.0 ps), and finally driven to 0 K with an even smaller time constant of 0.1 ps for another 5 ps. The time step during this process was 1 fs, and the SHAKE⁶³ algorithm was applied to all hydrogens in all our studies.

Next, we performed a temperature and pressure equilibration to 300 K and 1 atm, for 50 ps, using a Berendsen thermostat time constant of 1.5 ps and a weak coupling isotropic barostat with a relaxation time of 2 ps; the time step here was 2 fs. Finally, a 2 ns constant volume production run at 300 K (time step of 2 ps) was carried out, where the first 0.4 ns trajectory was used for equilibration. From the latest 1.6 ns, a frame was extracted every 40 ps, thus obtaining a sample of $n_s = 40$ frames. These frames were analyzed by our HSMD-TI procedure.

A similar procedure was applied to the solvent system consisting of L8 without the protein, which was solvated with 631 water molecules in a truncated octahedron. No counterions were added to this neutral system. Again, 40 frames were determined for this system as well. It should be pointed out that we have checked the movement of the ligand in the 2 ns MD trajectories using computer graphics. Rings 1 and 2 and the linear part of the peptide that starts from ring 3 (see Figure 1) were found to fully rotate in both the solvent and protein, where the ligand in solvent showed a random coil behavior. We therefore consider the 2 ns trajectories to be reasonably long for the present study, which is mostly concentrated on the development of HSMD-TI for various realistic models.

Table 1. Results for the Internal Entropy $S_{\text{ligand}}^{\text{A}}$ of the Ligand in the Solvent and Protein Environments and the Difference $\Delta S_{\text{ligand}}^{\text{A}} = S_{\text{ligand}}^{\text{A}}(\text{sol}) - S_{\text{ligand}}^{\text{A}}(\text{p})$ Obtained with AMBER99/PME^a

bin size, δ	n_f	$TS_{\text{ligand}}^{\text{A}}(\text{sol})$	$TS_{\text{ligand}}^{\text{A}}(\text{p})$	$T\Delta S_{\text{ligand}}^{\text{A}}$
$\Delta\alpha_k/30$	1000	3.2	0.8	2.4
	2000	-10.5	-12.7	2.2
	3000	-18.4	-20.2	1.8
	4000	-23.4	-25.3	1.9
	5000	-27.3	-29.2	1.9
	6000	-30.4	-32.3	1.9
$\Delta\alpha_k/60$	1000	3.2	0.9	2.3
	2000	-10.7	-12.9	2.2
	3000	-18.5	-20.3	1.8
	4000	-23.5	-25.4	1.9
	5000	-27.5	-29.3	1.8
	6000	-30.7	-32.4	1.7
$\Delta\alpha_k/90$	1000	3.1	0.8	2.3
	2000	-10.7	-12.8	2.1
	3000	-18.5	-20.3	1.8
	4000	-23.6	-25.5	1.9
	5000	-27.8	-29.5	1.7
	6000	-30.9	-32.6	1.7

converged

1.7 ± 0.2

^a The results were obtained by reconstructing $n_s = 40$ structures of L8 selected homogeneously from MD samples of 1.6 ns for the solvent and in the protein environments. The results are calculated as functions of $\delta = \Delta\alpha_k/l$ and n_f (eq 4)—the bin and sample size of the future chains, respectively. $S_{\text{ligand}}^{\text{A}}$ is defined up to an additive constant that is expected to be the same for both environments. The (best) results for $n_f = 6000$ are bold-faced.

III.3. Calculation of Entropy. L8 consists of three rings denoted 1–3 (Figures 1 and 2), where ring 3 is located deep in the active site as a basis for the two subchains which point to the outside of the pocket. Therefore, we defined the second carbon of ring 3 as the first reference atom (with coordinates \mathbf{x}_1 , see section II.5). This allows defining a physically meaningful external entropy which is related to the global movement of L8 in the active site. We first discuss results for the internal entropy. We reconstructed $n_s = 40$ L8 conformations (denoted i) in both, the solvent and the protein, where 31 atoms ($k' = 1, \dots, 31$) and 62 angles α_k ($1 \leq k \leq K = 62$) participate in the reconstruction (see section II.4). Each reconstruction step (out of 31) starts from conformation i with a 120 ps production run where a future L8 conformation is stored every 20 fs for a later analysis; thus, the total sample for each step consists of $n_f = 6000$ future conformations, where the first 200 are usually dropped as part of the equilibration. The number of counts, n_{visit} (eq 4), for each pair of bins is calculated, leading to $TP_{k'}$, where the product of the 31 TPs is the distribution, ρ^{HS} (eq 5), which leads to the entropy, S^{A} (eqs 6 and 7). This reconstruction is significantly larger than the 14-atom reconstruction performed previously for biotin.²²

In practice, the calculation is done in two stages, where in stage 1 we carry out the reconstruction simulations. Thus, for the 31 reconstructed atoms, $31n_f$ future chains are generated for snapshot i , and their coordinates are stored in a file for a later analysis in stage 2. Stage 1 can be performed in a straightforward way with any of the available programs, AMBER, TINKER, CHARMM,

etc. In stage 2, the files generated in stage 1 are read by an analysis program, which enables one to calculate the transition probabilities and to study the behavior of $S_{\text{ligand}}^{\text{A}}$ (eq 7), $\Delta S_{\text{ligand}}^{\text{A}}$ (eq 8), and S_{external} (eq 10) as a function of various parameters (e.g., bin size, n_s , and n_f) without the need to carry out additional (stage 1) runs. This free program with a tutorial and explanations appears at <http://www.ccbb.pitt.edu/Faculty/meirovitch/reconstruction-web/reconstruction-web.html>.

III.4. Results for the Internal Entropy. Results for $S_{\text{ligand}}^{\text{A}}$ (eqs 6 and 7) are presented in Table 1. It should be noted that the angles (α_k) are calculated in radians, which can lead to negative entropies (in contrast to our previous studies where using degrees led to positive entropies^{22,25–29}). This is not unexpected, as $S_{\text{ligand}}^{\text{A}}(\text{sol})$ is defined up to an additive constant, and we are interested only in entropy differences $\Delta S_{\text{ligand}}^{\text{A}}$ (eq 8) where this constant cancels out. As expected by theory, the results for $S_{\text{ligand}}^{\text{A}}$ decrease systematically as the approximation improves, i.e., with increasing n_f , but they remain unchanged as a function of δ . However, these results have not converged even for $n_f = 6000$. On the other hand, the corresponding results for ΔS^{A} show convergence to 1.7 ± 0.2 kcal/mol, which is thus considered to be the exact result. This relatively small difference might seem surprising, but as pointed out in section III.2, computer visualization shows that the two subchains (branches) of the ligand that protrude from the active site have significant conformational freedom, where rings 1 and 2 perform full rotations like in solvent.

III.5. Results for the External Entropy. The contributions of the three reference atoms to the external entropy appear in Table 2 as a function of n_f and only a single bin size $\Delta\alpha_k/60$, because exactly the same results (within the statistical errors) were obtained for bin sizes within the range $\Delta\alpha_k/30$ to $\Delta\alpha_k/90$. As expected, for each atom, the entropy decreases as the approximation improves, i.e., as n_f increases; however, the results have not completely converged even by increasing the maximal n_f to 12 000. The space covered by these variables can be estimated from eq 4, using the results obtained for the largest bin, $\Delta\alpha_k$, for which $n_{\text{visit}}/n_f = 1$, or more specifically by calculating $\exp(TS^{\text{A}}/0.6)$, where TS^{A} stands for the results in the table and $k_B T = 0.6$ at 300 K. Thus, atom 1 visits a volume of $\sim 14 \text{ \AA}^3$ where the “bond angle” and “dihedral angle” related to atom 2 cover together 7.2% of 4π ($=12.6$) and the range of change of the dihedral angle of atom 3 is $\sim 48^\circ$. However, the results for atom 2 and 3 should be considered as lower bounds since the predecessor atoms, 1 and 2, respectively are held fixed.

III.6. TI Results. To each of the $n_s = 40$ frames of the solvent sample we applied a TI procedure where the ligand–water interactions were turned off gradually for a fixed L8 structure; for the 40 frames of the protein environment, we decoupled both the ligand–water and ligand–protein interactions. Using a parameter λ , $0 \leq \lambda \leq 1$, the electrostatic interactions were decoupled first followed by decoupling the Lennard-Jones (LJ) potentials (in the presence of zero electrostatic interactions). In all, 30 λ values (windows) were used, 13 for the electrostatic interactions ($\lambda = 0.01, 0.05, 0.10, 0.20, 0.30, 0.40, 0.50, 0.60, 0.70, 0.80, 0.90, 0.95$, and 0.99) and 17 for LJ ($\lambda = 0.01, 0.03, 0.07, 0.10, 0.15, 0.20, 0.30, 0.40, 0.50, 0.60, 0.70, 0.80, 0.85, 0.90, 0.93, 0.97$, and 0.99). In the LJ integration, we used a soft-core potential based on $\delta = 3 \text{ \AA}$.^{29,64} For a given frame (i), each integration step (window) always started from the (initial) structure of i according to the corresponding step potential energy $[E(\lambda)]$, followed by a 120 ps production run, where the initial 20 ps are discarded for equilibration.

Table 2. External Entropy for the Three Reference Atoms Based on AMBER99/PME^a

bin size, δ	n_f	$TS_{\text{ligand}}^A(1)$	$TS_{\text{ligand}}^A(2)$	$TS_{\text{ligand}}^A(3)$	$TS_{\text{ligand}}^A(\text{total})$
$\Delta\alpha_k/60$	2000	1.3	-0.1	0.7	1.9
	4000	1.0	-0.6	0.2	0.6
	6000	0.8	-0.8	0.0	0.0
	8000	0.7	-0.9	-0.2	-0.4
	10000	0.6	-1.1	-0.3	-0.8
	12000	0.5	-1.2	-0.4	-1.1
$\Delta\alpha_k$	12000	1.6	-0.06	-0.1	1.4
space covered		14.4 Å ³	0.90 out of	48°	
			4 π = 12.6 (7.2%)		

^a The results are based on the 40 structures of L8 in the active site used to calculate the internal entropy (Table 1); they are presented for bin sizes $\delta = \Delta\alpha_k/60$ and $\Delta\alpha_k$ (eq 3) and n_f (eq 4)—the sample size of the future chains with a maximal value of 12000. $S_{\text{ligand}}^A(1)$ is the entropy related to the volume occupied by the first atom. $S_{\text{ligand}}^A(2)$ is related to the “bond angle” and “dihedral angle” of atom 2 and $S_{\text{ligand}}^A(3)$ to the dihedral angle of atom 3. In the last row, we provide the space covered by the variables of each atom. The statistical error for $-TS_{\text{ligand}}^A(\text{total})$ is 0.5 kcal/mol. The (best) results for $n_f = 12000$ are bold-faced, where $TS_{\text{ligand}}^A(\text{total}) = TS_{\text{external}}$.

The results for F^{TI} obtained by eliminating the charge (ch) and the LJ also include ligand–ligand contributions,⁶⁵ and because the conformations are fixed, these contributions constitute part of the ligand–ligand interaction energy. To calculate these energies, we generated two samples of the ligand in vacuum by removing the protein and water and applied TI to both (vacuum) samples. We have found the average result in solvent to be lower by 0.2 kcal/mol than the protein result. These very close results were subtracted from the original F^{TI} values, and the subtracted values are provided in Table 3, leading to $\Delta F^{\text{TI}} = F^{\text{TI}}(\text{p}) - F^{\text{TI}}(\text{s}) = -21.1$ kcal/mol (this result is expected to be converged as -20.6 was obtained for 60 ps runs). The ligand–ligand energy $E_{\text{ligand–ligand}}$ in Table 4 considers only the total bonded energy of the ligand, $E_{\text{bond}}(\text{p}) = 64.6$ and $E_{\text{bond}}(\text{s}) = 64.0$, i.e., $\Delta E_{\text{bond}} = 0.6$ kcal/mol.

Table 4 presents several energetic and entropic components that lead to our estimation of the absolute binding free energy, $\Delta A^0 = -10.7 \pm 1.0$ kcal/mol, which within the error bars is equal to the experimental values -10.9 (ref 37) and -10.6 (ref 66) kcal/mol. Notice, however, that the results for $-TS_{\text{external}}$ (Table 2) have not been converged, and we have estimated this uncertainty by specifying an increased statistical error of ± 0.5 kcal/mol to TS_{external} . This nonconvergent result is due to insufficient future sampling (too small n_f), whereas the number of frames studied, $n_s = 40$, is adequate. In fact, as discussed in sections II.3 and II.6, HSMD-TI requires a relatively small sample (number of frames). Indeed, the result for ΔF^{TI} (eq 12) is based on $n_s = 20$ frames, where $n_s \sim 15$ has led to the same result. While the entropy in principle requires larger samples than the free energy, converging results for ΔS_{ligand} have been obtained for $n_s \sim 40$.

Our result, $\Delta A^0 = -10.7 \pm 1.0$ kcal/mol, should be compared to results obtained by other methods. Pande’s group applied DAM to a set of eight ligands bound to FKBP using the AMBER and GAFF force fields, where the ligand–environment interactions were eliminated with FEP rather than TI. In their first paper,⁸ they obtained for L8 $\Delta A^0 = -7.3$ kcal/mol without

Table 3. AMBER99/PME Results for the Free Energy, F^{TI} in kcal/mol, Obtained by TI for the Protein and the Solvent Environments^a

	$F^{\text{TI}}(\text{ch})$	$F^{\text{TI}}(\text{LJ})$	F^{TI}
protein	-25.9 ± 0.4	-15.6 ± 0.9	-41.5 ± 0.7
solvent	-20.2 ± 0.2	-0.2 ± 0.2	-20.4 ± 0.2
$\Delta = \text{prot} - \text{sol}$	-5.7 ± 0.4	-15.4 ± 0.5	-21.1 ± 0.5

^a $F^{\text{TI}}(\text{ch})$ and $F^{\text{TI}}(\text{LJ})$ are free energies calculated by TI by eliminating the electrostatic and LJ interactions, respectively, based on $n_s = 20$ structures for each environment. F^{TI} is their sum.

considering the standard state contribution. Much better results have been obtained for these eight ligands in their following paper,¹³ where $\Delta A^0(\text{L8}) = -10.5 \pm 0.26$ kcal/mol. However, we feel that the active site volume, V^b , defined in ref 13 is too large depending strongly on the ligand size. Indeed, for all eight ligands, V^b is larger than V^0 where the maximal ratio is $V^b/V^0 \sim 2.8$.

In a third study by Fujitani et al.,¹² the force field was changed and the standard state correction was not applied; for L8, they obtained $\Delta A^0 = -10.1$ kcal/mol. Using DDM and the SSBP/GSBP modeling, Wang et al.⁴⁰ obtained $\Delta A^0 = -10.3$ kcal/mol with errors of ± 0.4 and ± 1.2 kcal/mol depending on the initial equilibration method. It should be pointed out that they used a flat bottom restraining potential on all symmetric units (rings) which might affect the entropy difference, ΔS (eq 8), hence the binding free energy. We also mention the results of Lamb et al.,⁴² who applied LIE⁶⁷ with various sets of parameters obtaining results for $\Delta A^0(\text{L8})$ between -10.1 and -10.3 kcal/mol. It should be noted that in our calculations none of the restrictions mentioned above have been imposed.

III.7. DDM Results Obtained with SSBP/GSBP. The use of PME requires treating the entire protein soaked in (typically) thousands of water molecules; hence for a large protein application of HSMD-TI (as well as other techniques), is expected to become time-consuming. This is the reason why we have studied initially the avidin–biotin complex with a model of partial structure, where only protein atoms and waters close to the active site were considered. However, the unsatisfactory results (discussed in section III.1) obtained for FKBP–L8 suggest that taking into account long-range electrostatic effects is imperative. Models of partial structure that take into account long-range electrostatics (the reaction field) have been developed, in particular by Warshel’s group (the program MOLARIS^{68–70}) and the SSBP/GSBP models mentioned earlier.^{9,43,44}

Therefore, we decided to apply HSMD-TI to the complex FKBP–L8 modeled by GSBP, where L8 in a solvent is modeled by SSBP; the relevant computer programs are included in the software package CHARMM.⁵⁰ While this system has already been studied by Wang et al.⁴⁰ using DDM, we have decided to apply DDM again where some of the parameters differ from those used by Wang et al. This would help us get a feel for the sensitivity of the model and establish a set of results that will be compared to later results obtained by HSMD-TI.

First, instead of using the GAFF parameters,⁴⁹ the force-field parameters for the ligand were taken from the CHARMM general force field,⁷¹ and the charges of the histidine residues were kept neutral. Most importantly, our DDM procedure is based on a single harmonic distance restraint with a force constant, $K_r = 7$ kcal/(molÅ²) (as compared to several different restrains applied by Wang et al.⁴⁰); finally, no flat-bottom restraining potential was imposed on all symmetric units (rings).⁴⁰

Table 4. Energetic, Entropic, and Free Energy Components (in kcal/mol) Which Contribute to the Absolute Free Energy of Binding, ΔA^0 (eq 12), Obtained for the Protein and Solvent with AMBER99/PME^a

	$k_B T \ln(8\pi^2 V^0)$	$-TS_{\text{external}}$	$-TS_{\text{ligand}}$	$E_{\text{ligand-ligand}}$	F^{TI}	total
protein	7.0	1.1 ± 0.5	32.3 ± 0.5	64.6 ± 0.3	-41.5 ± 0.7	63.2 ± 1.6
solvent			30.6 ± 0.6	64.0 ± 0.6	-20.4 ± 0.2	73.9 ± 1.3
$\Delta = \text{prot} - \text{solv}$	7.0	1.1 ± 0.5	1.7 ± 0.2	0.6 ± 0.7	-21.1 ± 0.5	-10.7 ± 1.0

^aThe table is organized according to eq 12. Results for TS_{ligand} , TS_{external} , and F^{TI} are taken from Tables 1, 2, and 3, respectively; most of the components are defined up to an additive constant, and only their difference has a physical meaning. The absolute free energy of binding is $\Delta A^0 = 10.7 \pm 1.0$ kcal/mol and is defined on the right-hand side of the bottom row.

Table 5. Free Energy Results in kcal/mol Obtained by Applying DDM to the SSBP/GSBP/CHARMM Model^{43,44a}

	$\Delta A(\text{FEP})_{\text{charge}}$	$\Delta A(\text{FEP})_{\text{LJ-attractive}}$	$\Delta A(\text{FEP})_{\text{LJ-repulsive}}$	ΔA_{R}	ΔA_{r}^0	total
solvent	15.0 ± 0.3	40.5 ± 0.2	-49.0 ± 0.2			6.2 ± 0.5
protein	14.7 ± 0.3	59.2 ± 0.2	-49.6 ± 0.5	1.5 ± 0.1	-5.0	20.8 ± 0.6
$\Delta = \text{solv} - \text{prot}$	0.1 ± 0.4	-18.7 ± 0.3	0.6 ± 0.5	-1.5 ± 0.1	5.0	-14.6 ± 0.8

^aThe elimination of the ligand-environment interactions was obtained by free energy perturbation (FEP) in three steps, treating first the electrostatic interactions then the attractive and repulsive LJ interactions. ΔA_{R} and ΔA_{r}^0 and $\Delta A^0 = -14.6$ kcal/mol are defined in eqs 13 and 14. For each environment, the results are averages of five FEP runs started from different ligand or protein–ligand structures.

The results which appear in Table 5 are averages of five FEP runs. Notice that with the GSBP/SSBP software the ligand–environment interactions are eliminated in the following order: electrostatic, attractive LJ, and repulsive LJ. The equations below are based on the notation of ref 5 (eqs 29–31), but to simplify the presentation, we define several more notations. Thus, $\Delta A_{\text{S}}(\text{FEP}) \equiv \Delta A_{\text{I}}^0$ (ref 5) and $\Delta A_{\text{P}}(\text{FEP})$ stand for the free energy change due to the elimination of the ligand–environment interactions in the solvent and the protein, respectively. ΔA_{R} is the free energy due to the application of the restraint, and ΔA_{r}^0 is the free energy due to the release of the restraint plus the effect of the standard state volume. Z is the partition function, $Z = 8\pi^2 V \bar{Z}$ (see eq 1)

$$\begin{aligned} \Delta A_{\text{I}}^0 &= -k_B T \ln \frac{Z_{\text{L},0} Z_{0,\text{N}}}{Z_{\text{L},\text{N}}} = \Delta A_{\text{S}}(\text{FEP}) \\ \Delta A_{\text{II}}^{*,0} &= -k_B T \ln \frac{Z_{\text{P} \dots \text{L},\text{N}}}{Z_{\text{P},\text{L},\text{N}}} = \Delta A_{\text{R}} + \Delta A_{\text{P}}(\text{FEP}) \\ \Delta A_{\text{r}}^0 &= -k_B T \ln \frac{V^0 Z_{\text{P},\text{N}} Z_{\text{L},0}}{V Z_{\text{P} \dots \text{L},\text{N}}} = -k_B T \ln \frac{V^0 K_{\text{r}}^{3/2}}{(2\pi k_B T)^{3/2}} \end{aligned} \quad (13)$$

where the free energy of binding ΔA^0 is

$$\Delta A^0 = \Delta A_{\text{I}}^0 - \Delta A_{\text{II}}^{*,0} - \Delta A_{\text{r}}^0 \quad (14)$$

The systems were simulated by Langevin dynamics at $T = 300$ K with a friction coefficient of 5 ps^{-1} . For each environment, five initial configurations were selected from the last 2 ns of a 4 ns run (where the first 2 ns were used for equilibration). The solvent system (described by SSBP) consists of 1000 TIP3P water molecules within a sphere of ~ 20 Å, whereas the protein system (described by GSBP) contains 325 waters within a sphere of ~ 15 Å. The elimination of the L8–water interactions was carried out in 20 windows for each type of interaction (electrostatic, attractive LJ, and repulsive LJ) where for each window the first 60 ps were used for equilibration and the next 120 ps for production. The

different FEP results appear in Table 5 where $\Delta A^0 = -14.4$ kcal/mol is expected to be a converged value, as $\Delta A^0 = -15.3$ kcal/mol has been obtained from an initial limited study based on 30 ps equilibration and 30 ps production (FEP) runs.

The fact that the binding free energy, -14.5 , is still lower by ~ 3.5 kcal/mol from the experimental value (hence from the very good result obtained by Wang et al.⁴⁰) can be attributed to the following: (1) Only a single (distance) restraint was applied which cannot restrict adequately the conformational freedom of the ligand in the active site. A better restriction was achieved by Wang et al. who used several restraints (among them an orientational restraint). (2) Wang et al. applied a flat-bottom restraining potential on all symmetric units (rings). (3) They used the GAFF parameters and probably charged histidines. However, another reason for the computational/experimental disagreement might be an inaccurate water density around the protein in our calculations. We have found that the results are sensitive to this density, which is difficult to apply accurately to the relatively small protein system. This is probably the reason why Roux and Deng have introduced a grand-canonical procedure to control the water density.⁷² Moreover, below, we demonstrate that application of HSMD-TI to the same model (FKBP–L8 described by SSBP/GSBP) leads to a ΔA^0 value comparable to that obtained above by DDM, which lends further support for the water density effect.

III.8. HSMD-FEP Results Obtained with SSBP/GSBP. We have also applied HSMD-TI to the SSBP/GSBP model, where the method actually becomes HSMD-FEP. We used the same parameters defined above for DDM, but studying for the perturbation 10 (rather than five) configurations for each environment. The main difference is that the elimination of the ligand–environment interactions has been applied to fixed ligand structures selected from the 2 ns initial Langevin runs of the protein and solvent systems discussed in section III.7.

The results in Table 6 for the internal entropy TS_{ligand} (eq 7) are not converged, but their difference $T\Delta S_{\text{ligand}}$ (eq 8) are converged nicely to 5.5 ± 0.2 kcal/mol, which is larger than the 1.7 kcal/mol obtained in the AMBER/PME calculations (Table 1). This suggests that the ligand in the active site is more restricted by

SSBP/GSBP than by AMBER/PME. This is also in accord with the smaller external entropy, E_{external} (eq 10), of -3.3 kcal/mol obtained in Table 7 than the -1.1 kcal/mol obtained in Table 2. Notice that, as in Table 2, the external entropy in Table 7 has not been converged, and an error of 0.5 kcal/mol has been assigned to it.

All of these results are summarized in Table 8. Notice that in contrast to Table 5, the results in Table 8 under the three columns of $\Delta A(\text{FEP})$ are the negative values obtained in the actual FEP (see the end of the paragraph following eq 11).

Table 6. HSMD Results for the Internal Entropy S_{ligand}^A of the Ligand in the Solvent and Protein Environments and the Difference $\Delta S_{\text{ligand}}^A = S_{\text{ligand}}^A(\text{sol}) - S_{\text{ligand}}^A(\text{p})$ Obtained for the SSBP/GSBP/CHARMM Model^{43,44a}

bin size, δ	n_f	$TS_{\text{ligand}}^A(\text{sol})$	$TS_{\text{ligand}}^A(\text{p})$	$T\Delta S_{\text{ligand}}^A$
$\Delta\alpha_k/60$	2000	-11.5	-15.9	4.4
	4000	-24.0	-29.1	5.1
	6000	-31.4	-36.7	5.3
	8000	-36.7	-42.0	5.3
	10000	-40.7	-46.2	5.5
converged				5.5 ± 0.2

^a The results were obtained by reconstructing $n_s = 20$ structures of L8 selected homogeneously from the Langevin dynamics trajectories generated for the results of Table 5. The results are shown for different sample sizes of the future chains, n_f (eq 4), but only for one bin size $\delta = \Delta\alpha_k/l$ where $l = 60$, because results for $l = 30$ and 90 are similar. S_{ligand}^A is defined up to an additive constant that is expected to be the same for both environments. The (best) results for $n_f = 10^4$ are bold-faced.

Table 7. External Entropy for the Three Reference Atoms Obtained for the SSBP/GSBP/CHARMM Model^a

bin size, δ	n_f	$TS_{\text{ligand}}^A(1)$	$TS_{\text{ligand}}^A(2)$	$TS_{\text{ligand}}^A(3)$	$TS_{\text{ligand}}^A(\text{total})$
$\Delta\alpha_k/60$	2000	1.0	-0.8	1.1	1.3
	4000	0.7	-1.2	-0.1	0.4
	6000	0.5	-1.5	-0.8	-1.8
	8000	0.3	-1.7	-1.4	-2.8
	10000	0.2	-1.8	-1.8	-3.4
$\Delta\alpha_k$	10000	1.5	-0.7	-0.2	0.6
space covered		12.2 \AA^3	0.3 out of	41°	
			$4\pi = 12.6$ (2.5%)		

^a The results are based on the 20 structures of L8 in the active site used to calculate the internal entropy (Table 6); they are presented for bin sizes $\delta = \Delta\alpha_k/60$ and $\Delta\alpha_k$ (eq 3) and n_f (eq 4)—the sample size of the future chains with a maximal value of 10^4 . For the meaning of $S_{\text{ligand}}^A(1)$, $S_{\text{ligand}}^A(2)$, and $S_{\text{ligand}}^A(3)$, see the caption of Table 2. In the last row, we provide the space covered by the variables of each atom. The statistical error for $-TS_{\text{ligand}}^A(\text{total})$ is 0.5 kcal/mol. The (best) results for $n_f = 10^4$ are bold-faced, where $TS_{\text{ligand}}^A(\text{total}) = TS_{\text{external}}$.

Table 8. Free Energy Results in kcal/mol Obtained by Applying HSMD-FEP to the SSBP/GSBP/CHARMM Model^a

	$k_B T \ln(8\pi^2 V^0)$	$-TS_{\text{external}}$	$-TS_{\text{ligand}}$	$\Delta A(\text{FEP})_{\text{charge}}$	$\Delta A(\text{FEP})_{\text{LJ-attractive}}$	$\Delta A(\text{FEP})_{\text{LJ-repulsive}}$	$E_{\text{ligand-ligand}}$	total
protein	7.0	3.4 ± 0.5	46.2	-17.7 ± 0.5	-62.5 ± 0.5	42.9 ± 0.8	189.2 ± 2.0	208.5 ± 1.3
solvent			40.7	-15.5 ± 0.1	-38.1 ± 0.3	47.4 ± 0.3	188.9 ± 1.9	223.4 ± 0.6
$\Delta = \text{prot} - \text{solv}$	7.0	3.4 ± 0.5	5.5 ± 0.2	-2.2 ± 0.8	-24.4 ± 1.0	-4.5 ± 1.0	0.3 ± 0.9	-14.9 ± 1.5

^a The elimination of the ligand-environment interactions was obtained by free energy perturbation (FEP) in three steps, treating first the electrostatic interactions, then the attractive and repulsive LJ interactions. $\Delta A^0 = -14.9 \pm 1.5$ kcal/mol is obtained by eq 12. For each environment, the FEP results are averages of 10 runs started from different structures, while the entropy was obtained in Tables 6 and 7 by reconstructing 20 structures. $E_{\text{ligand-ligand}}$ is the intraligand energy.

The table shows that the absolute free energy of solvation is $\Delta A^0 = -14.9 \pm 1.5$ kcal/mol, in accord with $\Delta A^0 = -14.6 \pm 0.8$ obtained in Table 5 using DDM, but with a larger error due to the uncertainty in the result for TS_{external} .

The efficiency of HSMD-TI(FEP) can be judged by comparing it to that of the DDM procedure applied to FKBP-L8 modeled by SSBP/GSBP.⁴⁰ First, each method is based on a pair of calculations where the ligand-environment interactions are annihilated (or created) in the protein and solvent environments using FEP, $n_s = 10$ and 5, such pairs were carried out by us and in ref 40, respectively. For each pair, two reconstructions are performed by HSMD, where in ref 40, three restraints were built and released by FEP (which requires six integrations in total). It is difficult to compare exactly the times involved in the reconstructions and for treating the restraints, as for example converging results for $\Delta S_{\text{ligand}}^A$ were obtained also for $n_s = 10$ and $n_f = 6000$ (rather than $n_s = 20$ and 10 000, see Table 6). Therefore, one can roughly say that DDM and HSMD-FEP have comparable efficiency; however, HSMD-FEP provides more information, the entropies $\Delta S_{\text{ligand}}^A$ and S_{external} .

III.9. Summary and Conclusions. HSMD-TI is a new method for calculating the absolute free energy of binding which does not suffer from the end point problem and is independent of DAM and DDM. It is of interest to view HSMD-TI from the perspective of DDM. Thus, to apply DDM efficiently, one seeks to limit the conformational freedom of the ligand in the active site by imposing various restraints, the stronger the restraints, the longer the time required for building them up from zero by TI (or FEP). With HSMD-TI, each ligand structure studied is fixed, and restraints are not needed. However, eliminating (by TI) the ligand-environment interactions for a fixed structure leads to some entropy loss (as compared to DAM and DDM), which is recovered by the reconstruction process. Thus, unlike DDM and DAM, HSMD-TI provides the difference in the internal entropy of the ligand in the two environments; one also calculates the external entropy, which constitutes an unbiased measure for the global movement of the bound ligand, providing thereby estimation for the size of the active site. Finally, the fact that HSMD-TI leads to ΔA^0 not as a result of two integrations (protein and solvent) but as a sum of entropic, energetic, and free energy components, enables one to gain a more complete picture of the various forces that determine the complex stability. Thus, HSMD-TI provides deep microscopic insights into the binding mechanism which are important from the academic point of view as well as for rational drug design.

In this paper, the scope of the theory has been extended, where we elaborate on the correlation between free energy fluctuations and the sample size, n_s . We have provided a more complete description of the reconstruction of the internal coordinates, pointing out the freedom in determining the first reference atom, and the order of the treated atoms. Still, one would seek to select

a first reference atom which leads to a physically meaningful external entropy that adequately expresses the global movement of the ligand in the active site. The reconstruction of the external coordinates has been described in detail, as well as the incorporation of bond stretching within the framework of the reconstruction process; potential situations where this latter effect should be considered have been discussed.

The unsatisfactory (preliminary) HSMD-TI results obtained by applying the AMBER99–GAFF–TIP3P potentials to the FKBP–L8 complex described by a *finite* model of partial structure (i.e., a model which is not based on periodic boundary conditions) suggest that long-range electrostatic effects cannot be ignored. Indeed, the excellent result for ΔA^0 obtained in this paper demonstrates the importance of long-range electrostatics, the effectiveness of PME, and the high performance of HSMD-TI. Also, as discussed in sections II.3, II.6, and III.6, HSMD-TI requires relatively small sample sizes.

Finally, we have also tested the SSBP/GSBP model,^{43,44} which takes into account long-range electrostatics and as a model of partial structure has the potential to be computationally less demanding than PME. We applied both DDM and HSMD-FEP to FKBP–L8 modeled by SSBP/GSBP and found comparable results for ΔA^0 , which, however, are slightly lower than the experimental value. This disagreement has been attributed mainly to inaccurate water density around the protein in our calculations. To develop HSMD-TI further, we plan to apply it in the next step to the complex FKBP12–FKS06, where the ligand FKS06 is significantly larger than L8.

AUTHOR INFORMATION

Corresponding Author

*Phone: (412) 648-3338. E-mail: hagaim@pitt.edu.

ACKNOWLEDGMENT

This work was supported by NIH grant 2-R01 GM066090-4.

REFERENCES

- Hermans, J.; Shankar, S. *Isr. J. Chem.* **1986**, *27*, 225–227.
- Jorgensen, W. L.; Buckner, J. K.; Boudon, S.; Tirado-Rives, J. *J. Chem. Phys.* **1988**, *89*, 3742–3746.
- Miyamoto, S.; Kollman, P. A. *Proteins* **1993**, *16*, 226–245.
- Gilson, M. K.; Given, J. A.; Bush, B. L.; McCammon, J. A. *Biophys. J.* **1997**, *72*, 1047–1069.
- Boresch, S.; Tettinger, F.; Leitgeb, M.; Karplus, M. *J. Phys. Chem. B* **2003**, *107*, 9535–9551.
- Zhou, H.-X.; Gilson, M. K. *Chem. Rev.* **2009**, *109*, 4092–4107.
- 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.
- Fujitani, H.; Tanida, Y.; Ito, M.; Jayachandran, G.; Snow, C. D.; Shirts, M. R.; Sorin, E. J.; Pande, V. S. *J. Chem. Phys.* **2005**, *123*, 084108–5.
- Deng, Y.; Roux, B. *J. Phys. Chem. B* **2009**, *113*, 2234–2246.
- Singh, N.; Warshel, A. *Proteins* **2010**, *78*, 1724–1735.
- Singh, N.; Warshel, A. *Proteins* **2010**, *78*, 1705–1723.
- Fujitani, H.; Tanida, Y.; Matsuura, A. *Phys. Rev. E* **2009**, *79*, 021914–12.
- Jayachandran, G.; Shirts, M. R.; Park, S.; Pande, V. S. *J. Chem. Phys.* **2006**, *125*, 084901–12.
- Hamelberg, D.; McCammon, J. A. *J. Am. Chem. Soc.* **2004**, *126*, 7683–7689.
- Mobley, D. L.; Chodera, J. D.; Dill, K. A. *J. Chem. Theory Comput.* **2007**, *3*, 1231–1235.
- Pohorille, A.; Jarzynski, C.; Chipot, C. *J. Phys. Chem. B* **2010**, *114*, 10235–10253.
- Mobley, D. L.; Dill, K. A. *Structure* **2009**, *17*, 489–498.
- Roux, B.; Nina, M.; Pomes, R.; Smith, J.-C. *Biophys. J.* **1996**, *71*, 670–681.
- Hermans, J.; Wang, L. *J. Am. Chem. Soc.* **1997**, *119*, 2707–2714.
- Deng, Y.; Roux, B. *J. Chem. Theory Comput.* **2006**, *2*, 1255–1273.
- Chen, P.; Kuyucak, S. *Biophys. J.* **2009**, *96*, 2577–2588.
- General, I. J.; Dragomirova, R.; Meirovitch, H. *J. Phys. Chem. B* **2011**, *115*, 168–175.
- White, R. P.; Meirovitch, H. *J. Chem. Phys.* **2004**, *121*, 10889–10904.
- White, R. P.; Meirovitch, H. *J. Chem. Phys.* **2005**, *123*, 214908–11.
- Cheluvvaraja, S.; Meirovitch, H. *J. Chem. Phys.* **2005**, *122*, 054903–14.
- Cheluvvaraja, S.; Meirovitch, H. *J. Chem. Theory Comput.* **2008**, *4*, 192–208.
- Cheluvvaraja, S.; Mihailescu, M.; Meirovitch, H. *J. Phys. Chem. B* **2008**, *112*, 9512–9522.
- Mihailescu, M.; Meirovitch, H. *J. Phys. Chem. B* **2009**, *113*, 7950–7964.
- General, I. J.; Meirovitch, H. *J. Chem. Phys.* **2011**, *134*, 025104–17.
- Meirovitch, H. *J. Mol. Recognit.* **2010**, *23*, 153–172.
- Ponder, J. W. *TINKER – Software Tools for Molecular Design*, version 5.0; Washington University School of Medicine: St. Louis, MO, 2009.
- Harrison, R. K.; Stein, R. L. *J. Am. Chem. Soc.* **1992**, *114*, 3464–3471.
- Schreiber, S. L. *Science* **1991**, *251*, 283–287.
- Kissinger, C. R.; Parge, H. E.; Knighton, D. R.; Lewis, C. T.; Pelletier, L. A.; Tempczyk, A.; Kalish, V. J.; Tucker, K. D.; Showalter, R. E.; Moomaw, E. W.; Gastinel, L. N.; Habuka, N.; Chen, X. H.; Maldonado, F.; Barker, J. E.; Bacquet, R.; Villafranca, J. E. *Nature* **1995**, *378*, 641–644.
- Griffith, J. P.; Kim, J. L.; Kim, E. E.; Sintchak, M. D.; Thomson, J. A.; Fitzgibbon, M. J.; Fleming, M. A.; Caron, P. R.; Hsiao, K.; Navia, M. A. *Cell* **1995**, *82*, 507–522.
- Van Duyne, G. D.; Standaert, R. F.; Karplus, P. A.; Schreiber, S. L.; Clardy, J. *Science* **1991**, *252*, 839–842.
- Holt, D. A.; Luengo, J. I.; Yamashita, D. S.; Oh, H.; Konialian, A. L.; Yen, H.; Rozamus, L. W.; Brandt, M.; Bossard, M. J.; Levy, M. A.; Eggleston, D. S.; Liang, J.; Schultz, L. W.; Stout, T. J.; Clardy, J. *J. Am. Chem. Soc.* **1993**, *115*, 9925–9938.
- Wilson, K. P.; Yamashita, M. M.; Sintchak, M. D.; Rotstein, S. H.; Murcko, M. A.; Boger, J.; Thomson, J. A.; Fitzgibbon, M. J.; Black, J. R.; Navia, M. A. *Acta Crystallogr., Sect. D* **1995**, *51*, 511–521.
- Shirts, M. R.; Mobley, D. L.; Chodera, J. D.; Pande, V. S. *J. Phys. Chem. B* **2007**, *111*, 13052–13063.
- Wang, J.; Deng, Y.; Roux, B. *Biophys. J.* **2006**, *91*, 2798–2814.
- Swanson, J. M. J.; Henschman, R. H.; McCammon, J. A. *Biophys. J.* **2004**, *86*, 67–74.
- Lamb, M. L.; Tirado-Rives, J.; Jorgensen, W. L. *Bioorg. Med. Chem.* **1999**, *7*, 851–860.
- Beglov, D.; Roux, B. *J. Chem. Phys.* **1994**, *100*, 9050–9063.
- Im, W.; Bernèche, S.; Roux, B. *J. Chem. Phys.* **2001**, *114*, 2924–2937.
- Darden, T. A.; York, D. M.; Pedersen, L. G. *J. Chem. Phys.* **1993**, *98*, 10089–92.
- 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.; Wang, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Kolossváry, I.; Wong, K. F.; Paesani, F.; Vanicek, J.; Liu, J.; Wu, X.; Brozell, S. R.; Steinbrecher, T.; Gohlke, H.; Cai, Q.; Ye, X.; Wang, J.; J.; Hsieh, M.-J.; Cui, G.; Roe, D. R.; Mathews, D. H.; Seetin, M. G.; Sagui, C.; Babin, V.; Luchko, T.; Gusarov, S.; A. Kovalenko, A.; Kollman, P. A. *AMBER 11*; University of California: San Francisco, CA, 2010.
- Cornell, W. D.; Cieplak, P.; Bayly, C. L.; Gould, I. R.; Merz, K. M., Jr.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. *J. Am. Chem. Soc.* **1995**, *117*, 5179–5197.

- (48) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926–935.
- (49) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. *J. Comput. Chem.* **2004**, *25*, 1157–1174.
- (50) MacKerell, A. D., Jr.; Bashford, D.; Bellott, M.; Dumbrack, 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.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wio2rkiewicz-Kuczera, J.; Yin, D.; Karplus, M. *J. Phys. Chem. B* **1998**, *102*, 3586–3616.
- (51) Alder, B. J.; Wainwright, T. E. *J. Chem. Phys.* **1959**, *31*, 459–466.
- (52) McCammon, J. A.; Gelin, B. R.; Karplus, M. *Nature* **1977**, *267*, 585–590.
- (53) Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. *J. Chem. Phys.* **1953**, *21*, 1087–1092.
- (54) Meirovitch, H. *J. Chem. Phys.* **1988**, *89*, 2514–2522.
- (55) The practical meaning of an exact method (such as Metropolis Monte Carlo⁵³) is that for a given statistical error Δ around the correct value of a property E , there is a sample size n_Δ in which for $n > n_\Delta$ the corresponding error $\Delta_n < \Delta$; thus, practically, the values E_n for $n > n_\Delta$ will converge, i.e., will be equal within the error Δ (one should use this criterion cautiously for systems with rugged conformational space). This criterion also applies to HSMD, where the accuracy of $\ln P_i$ for a ligand depends only on the reconstruction sample size n_r , and in principle a desired accuracy can be obtained by increasing n_r adequately. While in most cases obtaining convergence for $\ln P_i$ will be too time-consuming, entropy differences (i.e., in $\langle \ln P_i \rangle$ for the ligand in the protein and solvent environments) converge rapidly.
- (56) Meirovitch, H.; Alexandrowicz, Z. *J. Stat. Phys.* **1976**, *15*, 123–127.
- (57) Meirovitch, H. *J. Chem. Phys.* **1999**, *111*, 7215–7224.
- (58) Hnizdo, V.; Darian, E.; Fedorowicz, A.; Demchuk, E.; Li, S.; Singh, H. *J. Comput. Chem.* **2007**, *28*, 655–668.
- (59) Killian, B. J.; Kravitz, J. Y.; Gilson, M. K. *J. Chem. Phys.* **2007**, *127*, 024107–16.
- (60) Hnizdo, V.; Tan, J.; Killian, B. J.; Gilson, M. K. *J. Comput. Chem.* **2008**, *29*, 1605–1614.
- (61) We distinguish here between a physical time related to the original 2 ns trajectory from which the n_s frames were extracted and a procedural time which is related to the step-by-step reconstruction procedure of a single frame. The words future and past refer to the procedural time, where in the reconstruction of atom k' the past atoms (1, 2, ..., $k' - 1$) are held fixed, while the future atoms $k', k' + 1, \dots, K/2$ and all of the water molecules are free to move in the MD simulation.
- (62) Wang, J.; Dixon, R.; Kollman, P. A. *Proteins* **1999**, *34*, 69–81.
- (63) Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids*; Clarendon Press: Oxford, U. K., 1987.
- (64) Zacharias, M.; Straatsma, T. P.; McCammon, J. A. *J. Chem. Phys.* **1994**, *100*, 9025–9031.
- (65) Steinbrecher, T.; Mobley, D. L.; Case, A. C. *J. Chem. Phys.* **2007**, *127*, 21410813.
- (66) Hamilton, G.; Steiner, J. *Curr. Pharm. Des.* **1997**, *3*, 405–428.
- (67) Åquist, J.; Medina, C.; Samuelsson, J. -E. *Protein Eng.* **1994**, *7*, 385–391.
- (68) Warshel, A.; Sharma, P. K.; Kato, M.; Parson, W. W. *Biochim. Biophys. Acta* **2006**, *1764*, 1647–1676.
- (69) King, E.; Warshel, A. *J. Chem. Phys.* **1989**, *91*, 3647–3661.
- (70) Lee, F. S.; Chu, Z. T.; Warshel, A. *J. Comput. Chem.* **1993**, *14*, 161–185.
- (71) Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I.; Mackerell, A. D., Jr. *J. Comput. Chem.* **2010**, *31*, 671–690.
- (72) Deng, Y.; Roux, B. *J. Chem. Phys.* **2008**, *128*, 115103–8.

A Hybrid All-Atom Structure-Based Model for Protein Folding and Large Scale Conformational Transitions

Ludovico Sutto, Ilaria Mereu, and Francesco Luigi Gervasio*

Spanish National Cancer Research Center (CNIO), Structural Biology and Biocomputing Programme, Melchor Fernandez Almagro, 3, E-28029 Madrid, Spain

ABSTRACT: Structure-based models are successful at conjugating the essence of the energy landscape theory of protein folding with an easy and efficient implementation. Recently, their realm expanded beyond a single protein structure, and structure-based potentials have been used profitably to widely study complex conformational transitions. Still, when dealing with structural rearrangements between two, or more, well-defined structures, an unbiased and transferable description of the local backbone and side chain interactions could be advantageous. Here, we propose an all-atom model that merges a classical force field description of these local interactions with a structure-based long-range potential that takes into account the different conformations. We first validate the model simulating and characterizing the folding reaction and the transition state of two well-known proteins: the villin headpiece and the SH3 domain. Then, we characterize the activation mechanism of the catalytic domain of c-Src kinase. Such a process involves the conformational rearrangement of a large loop and the swing of an α helix. The appearance of a stable intermediate state in the free energy landscape between the two conformational end points suggests the mechanism of the loop opening. The low computational cost of the model together with the satisfactory accuracy of the results make it a promising approach to studying conformational transitions in large protein systems.

1. INTRODUCTION

Molecular simulations of biological systems are nowadays routinely exploited to investigate complex phenomena with a high resolution in time and space. However, a large number of relevant processes involving proteins take place on a time scale which is difficult to access with fully solvated all-atom force field simulations on modern computers. Protein folding, large scale structural rearrangement, and allosteric regulation are important examples where free energy barriers hinder classical molecular dynamics sampling. To overcome these difficulties, different routes have been followed, either reducing the number of degrees of freedom by using coarse-grained models^{1–4} or by accelerating the dynamics with enhanced sampling methods.^{5,6}

Pure structure-based models, or $G\bar{o}$ -models,⁷ with only native interactions are well suited to the study of protein folding since their Hamiltonians are by construction minimally frustrated.² The energy landscape is smoothly funneled toward the native structure, and folding occurs quickly and in a cooperative fashion. A variety of flavors has been proposed where the actual physicochemical interactions are all condensed in an effective native potential combined with different levels of geometric coarse-graining.^{8–11} These models represent indeed the simplest implementation of the energy landscape theory of protein folding¹² and are able to reproduce many aspects of the protein folding process.² Their main tangible advantage is speed, as they are orders of magnitude faster than all-atom simulations, which translates into the ability to study larger systems and for longer times. But they are also extremely useful in testing our ability to catch the relevant descriptors of the system and to understand their physical meaning. However, they lack the transferability of all-atom and other coarse-grained force fields.

Besides protein folding, structure-based models have recently been used to elucidate peculiar fold switching behaviors,¹³ to

study large amplitude conformational transitions,^{14–18} and to reproduce the B-DNA base flipping process.¹⁹ In these cases, the contributions of two distinct structures must be merged in the Hamiltonian to reproduce the conformational transition. Different approaches have been proposed, ranging from simple summation of the two potential energies¹³ to exponential averaging of two energy functions^{14,17} or by solving a secular equation.^{15,18,19} These models have by construction two energy minima corresponding to the input structures and lead to two free energy basins separated by an adjustable barrier that can be tuned on experimental data. However, it is not clear whether or not these mixing approaches correctly reproduce conformations outside the two energy basins (e.g., the transition state ensemble). For instance, no distinctions are made between the local angular and dihedral terms and the long-range interactions upon merging. A possible solution would be to mix structure-based approaches with some of the transferable terms of the atomistic force fields. A similar approach has been tried with DNA,¹⁹ while in proteins $G\bar{o}$ -like potentials have only been used as an extra biasing term on top of the original all-atom force field to predict the effects of mutations²⁰ or to accelerate folding.²¹ Here, to provide a generally applicable and predictive structure-based approach which includes a transferable description of the local backbone and rotamer conformations, we merge the bonded, angular, and dihedral terms of an up-to-date all-atom semiempirical force field with a structure-based, long-range term. In so doing, we keep the advantages of a structure-based model where the long-range interactions can be set to include one or more structures while gaining at the same time a more correct description of the backbone and side chain rotamers and dynamics along the simulated trajectory.

Received: August 4, 2011

Published: October 24, 2011

To assess the validity of this hybrid model, we first use a single basin version to set the $G\bar{O}$ term energy parameter and reproduce the folding of two well-studied globular proteins with two different secondary structures: the three helix bundle chicken villin headpiece subdomain (HP35) and the β -sheet rich SH3 domain of human c-Src. Both of these proteins are known to fold fast and cooperatively and have been extensively characterized experimentally.^{22–29} Subsequently, we use the two-state model to simulate and characterize the large conformational transition taking place in the catalytic domain of c-Src tyrosine kinase upon activation. The role of this kinase in cancer³⁰ fostered an intense experimental characterization of its structure and dynamics.^{17,31–35} Indeed several crystal structures exist with different poses of the catalytic domain in its active and inactive states.^{34,36} The activation involves the crossing of a free energy barrier corresponding to the rearrangement of a 23-residue loop located near the catalytic site between the two lobes of the kinase. Due to its transient nature, such structural transition is difficult to characterize experimentally. At the same time, all-atom MD simulations require either conspicuous simulation time and resources or ad-hoc sampling techniques to accelerate the rate of barrier crossing. Indeed, the activations of c-Src and close Src family members have been studied only recently either using computational methods with a C_α multistate $G\bar{O}$ model¹⁷ or merging the information from multiple short molecular dynamics runs.^{31,32} Here, we use the proposed two-state model to investigate the activation transition mechanism with an all-atom resolution.

2. COMPUTATIONAL METHODS

2.1. Model. The model takes into account all non-hydrogen atoms of the proteins while the solvent is treated implicitly by the Langevin dynamics. The Hamiltonian can be summarized by the sum of two distinct contributions, $H = H_{\text{ff}} + H_{\text{sb}}$, where H_{ff} contains only the following terms of the AMBER99SB-ILDN force field:³⁷ two-body bonded potential, three-body angle potential, four-body proper and improper dihedrals, and 1–4 short-range van der Waals potentials (two-body term between atoms separated by three bonds). These terms account for the correct backbone and side chain rotamers and geometries and are fully described elsewhere.³⁸ The H_{sb} term is the two-body structure-based potential and has the form:

$$H_{\text{sb}} = \sum_{i < j}^{\text{nativecontacts}} \varepsilon_0 \left[\left(\frac{r_{ij}^0}{r_{ij}} \right)^{12} - 2 \left(\frac{r_{ij}^0}{r_{ij}} \right)^6 \right] + \sum_{i < j}^{\text{nonnativecontact}} \varepsilon_H \left(\frac{r_H}{r_{ij}} \right)^{12}$$

where the first sum is over all of the native pairs ij conferring to each native contact a Lennard-Jones potential with common depth ε_0 at the native distance r_{ij}^0 . The second term is a generic hardcore repulsion for non-native atom pairs to avoid atom overlap with constants set to $\varepsilon_H = 1.0$ kJ/mol and $r_H = 0.2$ nm. The current conformation is specified by the set of pair distances r_{ij} .

Usually, what sets the energy scale in pure $G\bar{O}$ models is the depth of the native potential. Here, we have the force field contribution with its semiempirical parameters already correctly tuned to mimic the correct interactions around 300 K; hence, the strength of the native interaction is not a free parameter anymore. On the contrary, we have to set ε_0 to correctly reproduce the balance between the long-range interactions and the local interactions. Too large values of ε_0 would drive the model toward a

pure $G\bar{O}$ model, making pointless the force field contributions. On the other hand, ε_0 must be large enough to effectively account for the nonbonded interactions. For this reason, a screening on a range of ε_0 values was performed with the aim of identifying a coupling strength so to provide, on one side, the desired balance between folding speed and correct geometry, along with folding temperatures close to their experimental values. This survey led to the choice of $\varepsilon_0 = 3.8$ kJ/mol (see the Appendix for the technical details).

We chose, for the sake of simplicity, to use the same ε_0 value for all atom–atom native pair interactions regardless of the atom type, the residue, or the protein considered. Nevertheless, since it is an all-atom model, part of the residue characteristics such as its rotational degrees of freedom and its steric hindrance are automatically taken into account, relaxing the necessity for a pair-specific potential.

In the case of the conformational transition of the catalytic domain of c-Src, the structure-based term H_{sb} is the sum over the native contacts of both the open and closed conformations of the activation loop (A-loop). In order for such an approach to lead to the desired two free energy minima, the set of native contacts of each structure must be largely mutually exclusive. Otherwise, a structural arrangement of the backbone and side chains satisfying both sets of contacts would lead to a single global energy minimum structure, instead of the two distinct minima. In our case, the cutoff defining a native contact is small (5 Å) compared to the movement of the activation loop that spans a distance of 25 Å. Hence, the activation switch must proceed by breaking the contacts of the starting conformation to make those of the final structure, avoiding passing through conformations in which both sets of contacts are formed. Finally, the free energy minima are equally populated if the conformational entropy and the energetic contributions of the two structures are similar. This holds in our case since both conformations are compact and thus share a reasonably similar conformational entropy. Moreover, the number of native interactions specific to the open and to the closed structure is almost the same ($C_{\text{open}} \approx C_{\text{closed}}$, see below). As a result, the potential energies of the two conformations are indeed very close (-2438 ± 176 kJ/mol for the open and -2593 ± 116 kJ/mol for the closed conformation calculated averaging the long-range contributions coming from 2000 conformations picked in their respective free energy basins).

2.2. Native Contact Definition. Following the work of Wu et al.,³⁹ given a reference structure, a native contact between two atoms i and j belonging to two different, not adjacent, residues is defined when two requirements are met: (i) the atoms are within a cutoff distance $d_{ij} < 5$ Å and (ii) there is no other atom k such that $d_{ik} < d_{ij}$ and $\hat{j}ik < 35^\circ$ or $d_{jk} < d_{ij}$ and $\hat{i}jk < 35^\circ$. This definition approximates the screening effect occurring when the interaction between two atoms is shadowed by the presence of a third atom in between. With such a definition the dependence of the number of native contacts as a function of the cutoff distance increases more smoothly and slowly than it would without screening effects, thus relaxing the dependence of the model on the cutoff value.

We found 259 contacts for the chicken villin headpiece HP35 (PDB code 1YRF) and 382 for the SH3 domain of c-Src (PDB code 1FMK). In the case of the catalytic domain of c-Src (residues Gln250:Glu524, in human c-Src numbering), the closed loop conformation (PDB code 2SRC) has a total of 2146 contacts. A total of 723 of these contacts involve at least one atom in the region of major structural change upon activation.

To be sure not to limit excessively the protein flexibility, besides the activation loop (residues Ala403:Pro425) we include in the definition of this region also part of the N-lobe comprising the α C-helix (residues Ala256:Arg318), as shown in the red and green regions of Figure 6. The remaining $C_{\text{KD}} = 1423$ contacts are essentially the same in both open and closed conformation and will be referred to as “kinase domain contacts” (shown in cyan in Figure 6). $C_{\text{closed}} = 438$ contacts out of the 723 transition related contacts are specific to the closed conformation while the remaining $C_{\text{shared}} = 285$ are present in both open and closed conformation, and in that case the native distance is chosen as the smallest contact distance between the two conformations. Finally, $C_{\text{open}} = 433$ contacts of the open conformation (PDB code 1Y57) involve at least one atom in the transition region and are unique to the open structure. As a result, the total number of native contacts for the c-Src two basins model is $C_{\text{tot}} = C_{\text{KD}} + C_{\text{closed}} + C_{\text{open}} + C_{\text{shared}} = 2579$.

2.3. Simulation Details and Performance. Langevin dynamics have been performed with the molecular dynamics package GROMACS 4⁴⁰ using an inverse friction constant of 1 ps and an integration time step of 2 fs. A 500 ns-long simulation has been performed for both chicken villin headpiece HP35 and the SH3 domain. In the case of HP35, 15 replicas in the 270–350 K temperature range have been used. For the sharper transition of the SH3 domain, 40 replicas in the 350–370 K range have been used. In both cases, the exchange between adjacent replicas was attempted every 1 ps. The parallel tempering⁴¹ technique and the long simulation time allow for a full convergence of the thermodynamic properties.

To simulate the activation transition of the catalytic domain of c-Src, we adopted the bias exchange metadynamics (BE)⁴² approach in which replicas subject to different metadynamics biasing potentials are exchanged between Langevin dynamics simulations performed at the same temperature. In our case, we used $T = 310$ K and three replicas: one subject to the metadynamics bias with the contact map of the open form of the A-loop as a collective variable, one with the contact map of the closed form of the A-loop, and a third white replica with no metadynamics bias. In so doing, the sampling of the white replica conformations converges very quickly to the correct equilibrium distribution, and the free energy surface can be easily calculated.

To perform the metadynamics runs we used the PLUMED⁴³ plug-in for GROMACS.

As a reference, in the case of the 452 atoms of SH3, the model performs 102 ns/day per core per replica on a modern computer cluster. The same system simulated with an unbiased standard force field with implicit solvent performs 37 ns/day, while when it is solvated in 9120 water molecules in a 284 nm³ cubic box it slows down to 2.2 ns/day. Obviously, since the Hamiltonians and the number of simulated particles are different, the direct comparison of these performances has to be taken with care. A more meaningful comparison could be made by measuring the overall simulation time needed to converge the observables of interest. In so doing, we can then appreciate that in a few days of calculations, with a standard computer cluster, our model is able to reproduce thousands of folding/unfolding events of a globular protein allowing for a validation of its thermodynamics and the structures of the transition state ensemble at atomic resolution. On the contrary, a few folding/unfolding events, with an unbiased all-atom force field approach, still requires nowadays months of calculations on specialized hardware.⁴⁴

Similarly, the conformational transition between two known conformers separated by a high free energy barrier also requires months of computer time with a standard force field even with biased dynamics,⁴⁵ while we can sample the transition of c-Src in a few days of computer time. For these cases, the gain in time of our model is thus roughly 2 orders of magnitude over standard unbiased all-atom force fields.

2.4. Analysis. In order to characterize the folding reaction of HP35 and SH3, several parameters are used. The heat capacity at constant volume $C_V(T)$ as a function of temperature T is calculated as $C_V(T) = \sigma_E^2/(k_B T^2)$ where k_B is Boltzmann's constant and σ_E are the fluctuations of the energy. The latter are calculated from the most likely density of state of the system obtained collecting the energies of the 15 runs at different temperatures and optimally combining them with the multiple histogram method.^{46,47} To measure the folding reaction cooperativity, use is made of Chan's parameter⁴⁸ κ_2 , defined as the ratio of the van't Hoff enthalpy to calorimetric enthalpy: $\kappa_2 = 2T_{\text{max}}(k_B C_p(T_{\text{max}}))^{1/2}/\Delta H_{\text{cal}}$ where T_{max} is the temperature of the C_p peak and ΔH_{cal} is the calorimetric enthalpy of the reaction determined as the integral of the heat capacity across the transition region. We used the energy and the C_V instead of the enthalpy and the C_p to calculate κ_2 since for our purposes they can be considered equivalent.

Another useful parameter to characterize folding is the native contact fraction Q , calculated as the ratio of the number of formed native contacts over the total number of native contacts. A native contact is considered formed if the atoms are within 1 Å of their native distance.

To characterize the folding transition state (TS) of HP35 and SH3, the Φ values per residue can be approximated by $\Phi_i = \langle Q_i \rangle_{\text{TS}}$ where Q_i is the fraction of formed native contacts of residue i and where the average $\langle \dots \rangle_{\text{TS}}$ is done over the set of conformations constituting the transition state.^{49,50}

Two structural parameters are also used to measure the similarity between a current structure Γ and a reference structure Γ^0 : the distance root-mean-square deviation $\text{dRMSD} = (1/(N(N-1)\sum_{i<j}(d_{ij}-d_{ij}^0)^2))^{1/2}$ and the root-mean-square deviation $\text{RMSD} = (1/N\sum_{i=1}^N(r_i-r_i^0)^2)^{1/2}$. Both quantities are calculated using N C α atoms and involve either their pairwise distance d_{ij}^0 or their position r_i^0 in the reference structure. In the latter case, r_i indicates the position of the C α atom i of the structure Γ aligned on the reference Γ^0 .

The free energy as a function of a generic reaction coordinate q is calculated from the equilibrium probability distribution $p(q)$ obtained during the simulation as $F(q) = -k_B T \ln(p(q))$.

3. RESULTS AND DISCUSSION

3.1. Folding of Villin Headpiece HP35 and SH3 Domain. In order to test the ability of our model to correctly fold both α -helical and β -sheet rich proteins, we chose two different domains: HP35 and SH3. HP35 consists of three α -helices packed together and, with its 35 residues, is one of the smallest proteins showing a cooperative folding.⁵¹ Several computational^{25,52–55} and experimental^{25–29,51,56} works thoroughly characterized its folding reaction describing the transition state as well as its secondary structure rich unfolded state. The chosen value of the native potential energy ϵ_0 results in a cooperative folding transition for HP35 with a folding temperature $T_f = 316$ K. Indeed, at $T = T_f$ the heat capacity C_V features a peak, and the average fraction of native contacts $\langle Q \rangle$ shows its steepest variation (Figure 1).

The fraction of native contacts Q is not sufficient alone to distinguish the role of each helix during the folding reaction. For this reason, the average fraction of native contact $\langle Q \rangle$ is decomposed according to four subsets of native contacts (Figure 1, lower panel), and the free energy is projected along both Q and the C_α distance root-mean-square deviation of the first and second helices ($dRMSD_{I-II}$) and the second and third helices ($dRMSD_{II-III}$) in Figure 2. The folding barrier $\Delta G_{U\ddagger} = 2$ kJ/mol

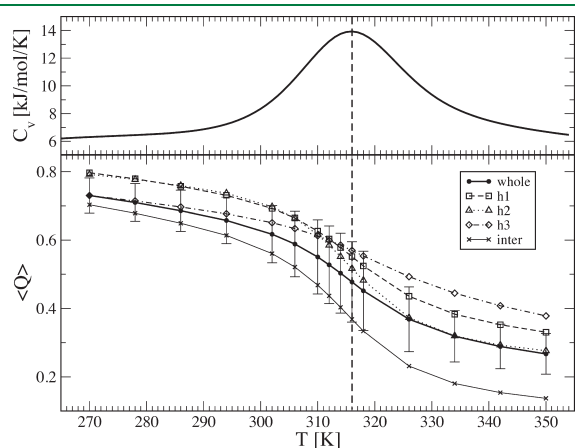


Figure 1. Thermodynamics of HP35 folding. In the upper panel, the heat capacity C_V is reported as a function of temperature T . Its peak at $T_f = 316$ K defines the folding temperature. In the lower panel, the average fraction of native contact as a function of the temperature of the whole protein (solid thick line, filled dots) is shown together with its standard deviation. In the same plot, the average fraction of native contacts divided into different subsets is also displayed: the internal contacts of helix 1 (dashed line, square), helix 2 (dotted line, triangles), and helix 3 (dash-dotted line, diamonds) and the tertiary native contacts between them (solid line, crosses). These curves highlight that the folding reaction mainly depends on the assembly of the helices while they keep a certain degree of nativeness even at high temperatures. As expected by its small size, helix 2 is the least stable and the one that loses most structure at high temperatures. On the contrary, helix 3 is the most stable and retains almost 40% of its structure at the highest sampled temperature.

at $T = 1.03T_f$ is relatively small because of the high content of residual secondary structure in the unfolded state. The conformations populating this unfolded basin are characterized by the loss of the majority of the tertiary native structure ($\langle Q_{\text{inter}} \rangle \approx 0.2$, Figure 1) and most of the helical content of helices I and II ($dRMSd_{I-II} \approx 0.4$ nm, Figure 2), while helix III maintains roughly half of its internal native interactions ($dRMSd_{II-III} \approx 0.2$ nm, Figure 2 and $\langle Q_{h3} \rangle \approx 0.5$, Figure 1). This residual structure, which involves the N-terminal residues of helix III, vanishes at $T = 1.05T_f$ leading to an extended unfolded basin (Figure 2, right-hand panels).

Despite the absence of any appreciable intermediate state with native secondary structure and poor tertiary interactions as found by solid-state NMR experiments,²⁷ the description of the unfolded state rich in helical content, in particular in the helix III region, is in good agreement with all-atom molecular dynamics calculations in explicit solvent^{52,55} as well as with recent triplet–triplet energy transfer experiments.^{25,26} A number of independent

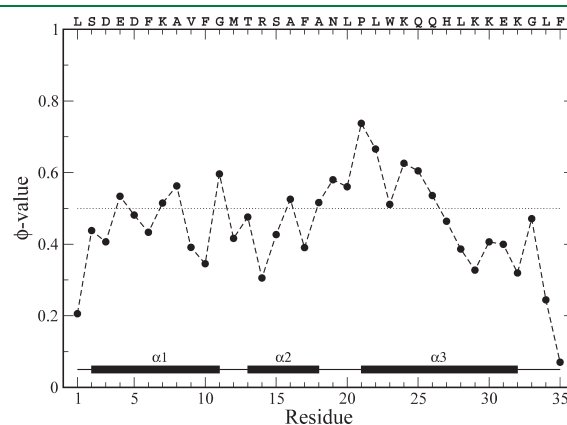


Figure 3. Characterization of HP35 folding transition state. The Φ value is calculated for each residue using a set of 5178 structures picked from the top of the folding free energy barrier at $T = T_f$. The highest values correspond to residues Pro21 and Leu22 belonging to the first turn of helix III. The helices are represented along the sequence as solid bars, and a dotted line is shown at $\Phi = 0.5$.

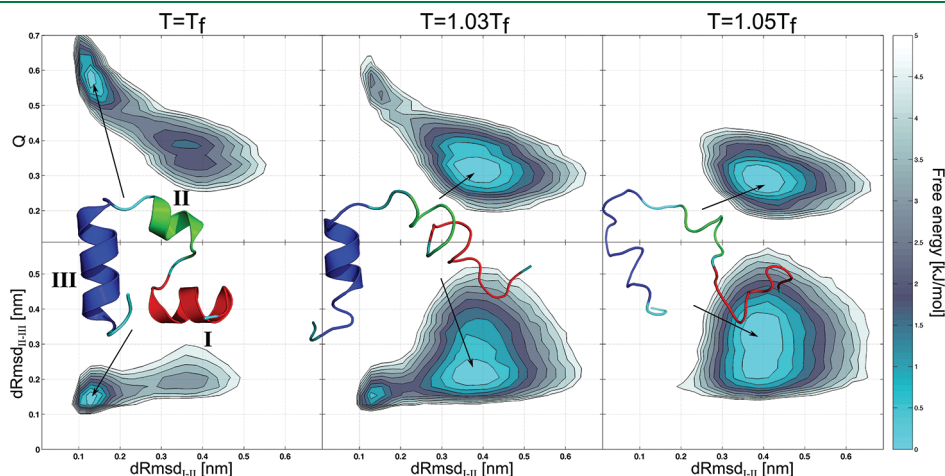


Figure 2. Free energy surfaces of HP35. The free energy surface at three increasing temperatures is projected along either the fraction of native contacts Q and the distance root-mean-square deviation with respect to the first and second helices ($dRMSD_{I-II}$) in the upper row or the distance root-mean-square deviation with respect to the second and third helices ($dRMSD_{II-III}$) and $dRMSD_{I-II}$ in the lower row. The contour lines are drawn every 0.5 kJ/mol. Three representative conformations of each basin are also shown.

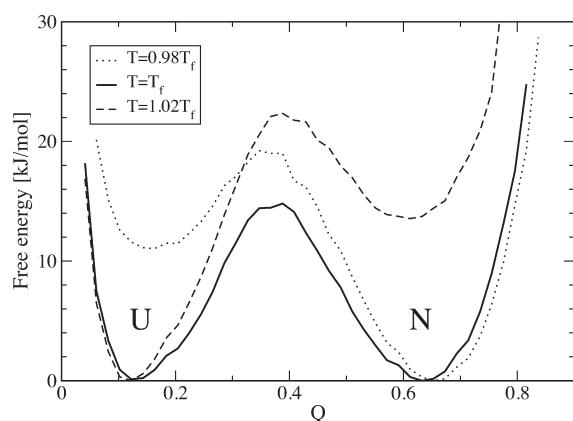


Figure 4. Free energy of the SH3 domain as a function of the fraction of formed native contacts Q at three different temperatures. At $T = T_f$ the native (N) and unfolded (U) states are equally populated (solid curve) and separated by a free energy barrier of about 14 kJ/mol located at $Q = 0.4$. At lower ($T = 0.98T_f$, dotted curve) or higher temperatures ($T = 1.02T_f$, dashed curve), the native or unfolded state is respectively predominant.

computational works also found an early formation of helices II and III, and a slower formation of helix I.^{53–55} Moreover, recent molecular dynamics simulations of the HP35 folding found the intermediate state caused by an incorrect docking of helices I and III where non-native interactions play an important stabilization role.^{25,52} This could explain the absence of the intermediate state in our calculations given the lack of any non-native tertiary interactions in our $G\bar{\sigma}$ -like potential.

The putative transition state, corresponding to the top of the free energy barrier, features the formation of most of the native contacts of Pro21 and neighboring residues as visible from the highest Φ value (Figure 3). This corresponds to the formation of the loop between helices II and III and the first turn of helix III. This residue is indeed known to play a crucial role for the protein stability since if mutated to alanine the villin headpiece is unable to fold to a native structure anymore.⁵⁷ The N-terminal half of helix III as well as residues 4, 7, and 8 belonging to helix I have more than 50% of the native interactions formed in the transition state ($\Phi > 0.5$). The last forming interactions, in the free energy descent toward the native state, involve the tail of helix III with both internal and tertiary contacts with the head of helix I that eventually lock the protein in its characteristic bundle.

To check the ability of our model to correctly fold also β -sheet structures, we simulated the 60-residue globular domain SH3 and repeated the thermodynamic analysis and folding characterization. As expected, the folding transition of the β domain SH3 is much more cooperative than for the α -helical HP35 as measured by Chan's parameter⁴⁸ κ_2 resulting in $\kappa_2^{\text{SH3}} = 0.80$ as compared to $\kappa_2^{\text{HP35}} = 0.53$ (see the Computational Methods). Intuitively, an α helix can form independently of the rest of the protein since it involves only local interactions, while a β sheet requires a wider structural organization. The three α helices of HP35 form gradually, losing entropy in favor of the energetic contributions of the local contacts. On the contrary, the formation of the β sheets of SH3 involves longer-range interactions with a larger entropic cost compensated by the formation of several native interactions, resulting in an emphasized all-or-none character.

To further support this picture, we calculated the difference in internal energy between the native and the unfolded state for

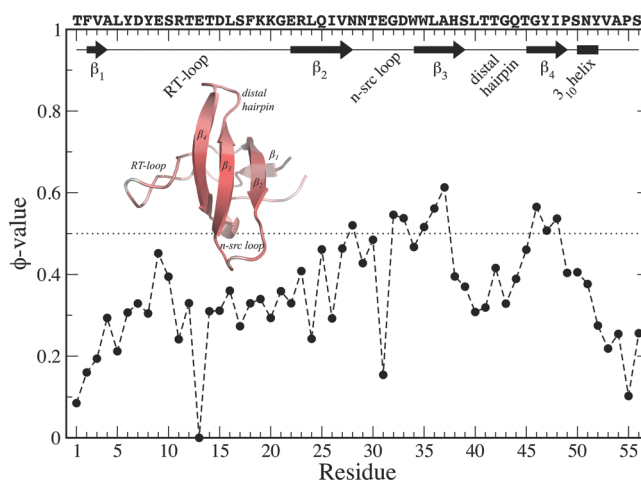


Figure 5. Characterization of SH3 folding transition state. Φ values are calculated for each residue using a set of 1350 structures picked from the top of the folding free energy barrier at $T = T_f$. Along the upper x axis is shown the sequence and a schematic representation of the secondary structure. A native conformation with color coded Φ values (darker color being 1, lighter color being 0) is also shown. A dotted line is plotted at $\Phi = 0.5$.

HP35 and SH3 at their respective folding temperatures. At $T = T_f$ we have by definition $\Delta G_{\text{UN}} = 0$ thus $T_f \Delta S_{\text{UN}} = \Delta E_{\text{UN}}$ where ΔG_{UN} , ΔS_{UN} , and ΔE_{UN} are respectively the difference in free energy, entropy, and internal energy between the unfolded (U) and native (N) states. Picking 30 000 structures in both basins for both proteins, the average difference in internal energy is $\Delta E_{\text{UN}}^{\text{HP35}} = 134 \pm 94$ kJ/mol and $\Delta E_{\text{UN}}^{\text{SH3}} = 647 \pm 129$ kJ/mol, accounting for the 6% and 25% of their respective average native state energy. As a result, the chicken villin shows a much lower difference in internal energy between structures of the unfolded and the native state due to the large amount of residual helical structure in the U state. On the contrary, the unfolded basin of SH3 is a minimum in free energy mainly because of its large entropy, which compensates the poor native content responsible for the large difference in internal energy between the U and N state conformations.

The thermodynamic stability of SH3 is also higher than for HP35 with a folding temperature of $T_f = 357$ K. The free energy projected along the fraction of native contacts Q shows two basins of equal depth at $T = T_f$ corresponding to the native (N) and unfolded (U) states separated by a high barrier of 14 kJ/mol (Figure 4).

The ensemble of structures on the top of this free energy barrier, containing about 40% of formed native contacts (Q in $[0.31–0.44]$), corresponds to the transition state of our model. These native contacts are not distributed homogeneously along the sequence but are rather concentrated in specific regions. Indeed, from the calculated Φ values, the most natively structured residues involve the n-src loop and the β_3 and β_4 strands (see Figure 5). Consequently, the less formed regions involve the RT-loop and the N- and C-terminal regions. These observations are in good agreement with the experimental characterization of the transition state done by the thorough Φ -value analysis of Riddle et al.²³ where they find the central three-stranded β sheet with the most structured residues. The linear correlation coefficient between the calculated Φ values Φ_{calcd} and the 37 available experimental ones Φ_{exptl} is $r = 0.48$ with a root-mean-square

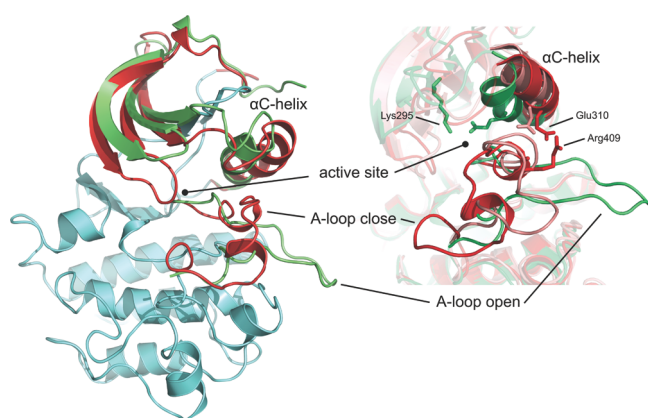


Figure 6. (Left) The catalytic domain of *c*-Src in its open (green) and closed (red) conformations. The cyan part, common to both conformations, defines the kinase domain native contacts while the open-specific contacts are taken from the green region, and the closed-specific native contacts from the red one. (Right) Close-up of the activation loop and the α C-helix from another viewpoint. Besides the open and closed conformations, also a representative conformation of the intermediate C^* basin is shown in pink. The catalytically critical Glu310 is also shown with sticks, together with the Lys295 and Tyr382 residues that stabilize the α C-in and α C-out conformations, respectively.

deviation of $\sigma = ((\sum (\Phi_{\text{exptl}} - \Phi_{\text{calcd}})^2 / N))^{1/2} = 0.32$. This result shows a semiquantitative agreement with the actual transition state similarly to other all-atom structure-based models^{58,59} while performing better than a C_α heterogeneous $G\bar{o}$ model.⁹ In particular, we overestimate the native content of Gly46 of strand β_4 while underestimating that of Gly43 belonging to the distal hairpin. In fact, when these two outliers are removed, the correlation coefficient of the remaining 35 Φ values increases to $r = 0.62$ with a smaller RMSD $\sigma = 0.29$.

These results validate the model as a valuable tool to reproduce the folding of both α and β protein domains. Indeed, the transition state ensembles are accessible with atomic detail and correspond with an acceptable confidence to the experimental ones, as shown by Φ -value analysis. This suggests that the model could be equally fit to study the large-scale conformational transitions in larger proteins. The main difference between folding and conformational transitions stems from the fact that, in the latter case, the conformations of the two extremes of the reaction are precise structures, while folding proceeds from a large ensemble of structures belonging to the unfolded state. Exploiting this difference through a two-basin energy landscape, we are able to characterize the activation of the catalytic domain of *c*-Src kinase.

3.2. The Activation Loop Dynamics of *c*-Src Catalytic Domain. *c*-Src is a nonreceptor tyrosine kinase critically involved in the regulation of fundamental cellular processes including cell growth, differentiation, and migration.⁶⁰ These functions are carried out through phosphorylation of the tyrosine residue of a protein substrate once the kinase is in its active conformation.^{34,60–62} An elevated and uncontrolled activation of *c*-Src is associated with many tumor types,^{30,62,63} highlighting the importance of the structural and dynamical characterization of its activation mechanism.

The catalytic domain is constituted by 274 residues forming a characteristic two-lobe structure with the catalytic active site located in a deep cleft between them (see Figure 6). The activation mechanism

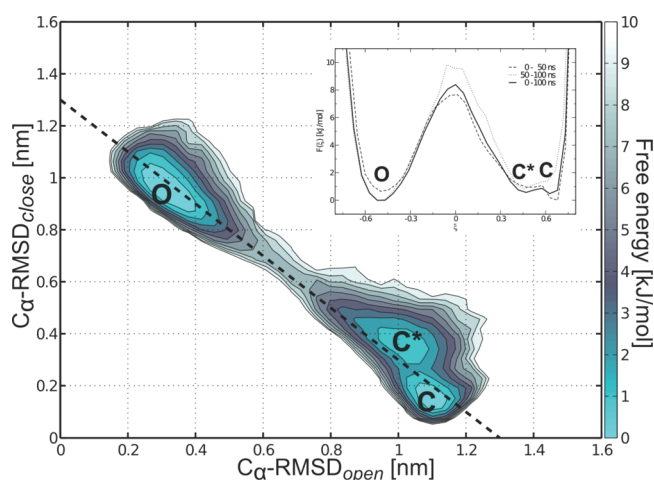


Figure 7. Transition free energy of *c*-Src at $T = 310$ K. The two main minima corresponding to the open (O) A-loop and to the closed (C) A-loop conformations are separated by a free energy barrier of 8 kJ/mol. A secondary shallow minimum in the closed basin corresponds to an intermediate state of conformations (C^*) where the activation loop adopts a particular hybrid structure. The contour lines are drawn every 1 kJ/mol. In the inset is shown the free energy projected along the dashed line of the bidimensional plot, corresponding to the linear combination of $\text{RMSD}_{\text{open}}$ and $\text{RMSD}_{\text{closed}}$ that best distinguishes the minima (solid line). In this monodimensional projection are also shown the free energies calculated for the first half of the simulation (dashed line) and for the second half (dotted line).

involves the rearrangement of a centrally located loop, called the activation loop or A-loop (residues Ala403 to Pro425), that switches from a folded and closed conformation obstructing the active site to a more extended one which serves as a platform for substrate binding.⁶¹ At the same time, the α C-helix (residues Pro304 to Lys315) in the N-terminal lobe also swings from an α C-out to an α C-in conformation. In so doing, the catalytically critical Glu310 points inward to the active site pocket where, together with the DFG motif (Asp404, Phe405, Gly406) at the beginning of the A-loop, it coordinates the transfer of the γ -phosphoryl group of ATP to the tyrosine residue of a protein substrate.⁶⁴ The open A-loop conformation is stabilized by the phosphorylation of a tyrosine residue (Tyr416), located in the middle of the loop, albeit the open conformation has been crystallized also with the unphosphorylated tyrosine.³⁶ This tyrosine is in fact accessible for phosphorylation only when exposed to the solvent and thus when the loop is not in its fully closed conformation. This suggests that the open and closed conformations are both accessible to the unphosphorylated catalytic domain, in agreement with the significant catalytic activity retained by the unphosphorylated form.^{65,66} As a consequence, we did not consider any phosphorylation effect in our model, and we neglected the two regulatory domains SH2 and SH3, focusing our attention on the sole ability of the catalytic domain to undergo the conformational rearrangement.

To characterize the conformational switch, a bidimensional free energy has been calculated from the simulation. The parameter that better distinguishes the structural change is the root-mean-square deviation (RMSD) of carbon α atoms of the A-loop with respect to either the open or closed conformation after proper alignment of the whole catalytic domain to the respective reference structure. In such space, the projection of the free energy shows two major basins, C and O, separated by a barrier of 8–9 kJ/mol: one corresponding to the folded, closed (C), A-loop, while the other corresponds to the extended, open (O),

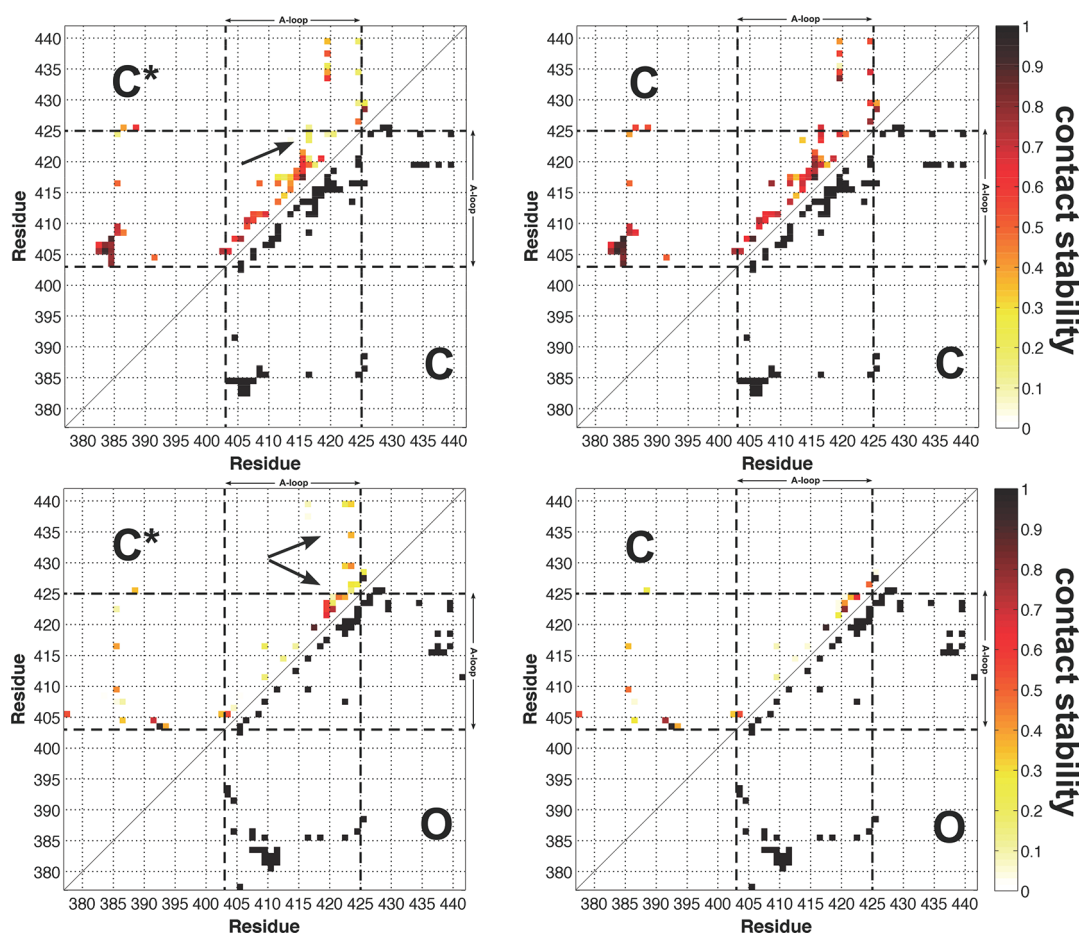


Figure 8. Stability contact maps of the A-loop of the intermediate state (C^*) compared to the closed state (C). The represented quantity is the probability of the native contact formation per residue calculated out of a set of 2000 conformations picked from either the intermediate state (C^* , left-hand side) or the closed state (C , right-hand side). For an easier reading, the native contacts are divided into closed-specific native contacts (top row) and open-specific native contacts (bottom row). The intermediate state contains hybrid conformations in which the C-terminal region of the activation loop partly disrupts its closed-like contacts in favor of open-like ones, as highlighted by the arrows.

conformation (see Figure 7). In the 100 ns simulation run, the catalytic domain undergoes multiple barrier crossings helped by the metadynamics bias on two of the three replicas, guaranteeing the convergence of the free energy profile (see the Computational Methods). To assess such convergence, the free energies of the first and second halves of the trajectory have been projected on a linear combination of $\text{RMSD}_{\text{open}}$ and $\text{RMSD}_{\text{closed}}$ for a more convenient comparison (see inset of Figure 7). The variable ξ , defined as $\xi = \text{RMSD}_{\text{open}}/(2)^{1/2} - \text{RMSD}_{\text{closed}}/(2)^{1/2}$, shows a difference of 2.5 kJ/mol $\approx 1k_B T$ in the top of the barrier as the largest deviation between the first and second half of the simulation, providing an estimate of the error in the free energy determination.

Interestingly, adjacent to the closed free energy basin C , a shallower second basin emerges in the bidimensional FES. It is separated by a small barrier to state C but nonetheless populated by well characterized structures. Such a free energy minimum, hereby denoted as C^* , corresponds to an intermediate state between the closed and open conformations that shares specific native contacts of the open loop and the closed loop. More precisely, C^* is populated by an ensemble of conformations where the centrally located residues of the A-loop (residues 411–419 which comprise most of a short 1.5 turn helix) form less stable interactions with their neighbors than in the closed

basin C (see a representative conformation in Figure 6 and the contact maps in the top row of Figure 8). This is even more pronounced in the region near the C-terminal end of the A-loop where the probability of closed-specific native contact formation between residues 419–425 and the following residues of the C-lobe, residues 426–439, falls below 0.2. At the same time, a cluster of open-like native interactions appears at the C-terminal end of the A-loop (see Figure 8, bottom row), while they are totally absent in the C state.

These observations suggest that the switching mechanism of the activation loop passes through an intermediate state (C^*) characterized by loose short-range interactions of the C-terminal half of the loop where a hybrid conformation of the activation loop is stably populated. This intermediate state has more open-like content than the C state, mostly located in the C-terminal end of the loop, indicating that this region is most prone to starting the transition to the full open structure (see Figure 6).

Besides the rearrangement of the A-loop, the activation of the catalytic domain also involves the rotation of the αC -helix where the catalytically critical Glu310 residue breaks its contacts with Tyr382, Arg409, and Leu410 and points toward the active site. In the active site pocket, favorable interactions between Glu310 and Lys295, Ile336, and the DFG-motif stabilize the αC -in conformation. Analyzing the fraction of native contacts made both by

Table 1. Inspection of the A-loop Closed (C), Intermediate (C*), and Open (O) State in Terms of the Conformation of the α C-helix^a

	closed (C)	intermediate (C*)	open (O)
$\langle Q_{\alpha C-in} \rangle$	0.51 (0.37)	0.52 (0.39)	0.57 (0.50)
$\langle Q_{\alpha C-out} \rangle$	0.53 (0.43)	0.53 (0.42)	0.43 (0.29)

^aThe fraction of formed native contacts specific to the α C-helix in (out) conformation $Q_{\alpha C-in}$ ($Q_{\alpha C-out}$) is averaged over a set of 2000, 2000, and 6000 structures from the C, C*, and O ensembles, respectively. In parentheses, only a subset of the native contacts of the α C-helix is considered, namely, those involving the Glu310 residue. The statistical standard deviation for all of the calculated quantities is around 0.07.

the α C-helix and also specifically by Glu310, we do not find a sharp correlation between the rotation of the helix and the conformation of the A-loop. In other words, in the ensemble of both open and closed conformations, we can find structures with either an inward rotation of the α C-helix or an outward rotation albeit with slightly different propensities (see Table 1). More in general, the helix rearrangement occurs with small cooperativity through a set of conformations stabilized by a blend of favorable interactions coming from both the α C-in and α C-out reference structures. In fact, in the closed and intermediate basins (C and C*), the α C-helix forms roughly half of the native contacts specific to the α C-in and half of those specific to the α C-out conformation, while the structures found in the open basin (O) are mainly characterized by the expected α C-in conformation (see Table 1).

Taken together these results are suggestive of the existence of an intermediate step in the activation transition of c-Src and the ability of the α C-helix to swing between two distinct conformations. These features have been previously predicted by short unbiased all-atom runs along an optimal free energy activation pathway.³¹ In that case, the intermediate state was structurally different from the present C* ensemble of conformations showing a clearer open-like structure of the A-loop and an outward rotation of the α C-helix. Interestingly, targeted molecular dynamics (TMD) simulations of the Src family member Lyn suggested a first barrier to the activation to be the helix–coil transition of the 1.5 turn helix at the C-terminal end of the A-loop.⁶⁷ This short helix involves residues that we find to be indeed much less structured in the proposed C* intermediate state. Further evidence of the stepwise mechanism of activation comes from the Src family member Hck where two distinct intermediate states have been observed in the activation transition together with an alternative activation pathway involving the partial unfolding of the N-lobe.³² An intermediate state has also been found in CDKs, Abl, and EGFR tyrosine kinases using either metadynamics molecular dynamics⁴⁵ or TMD.⁶⁸

While most features predicted by our hybrid model are in accordance with several computational and experimental observations, the predicted peculiar behavior of the α C-helix was not reported previously either in c-Src or in the close relatives Hck and Lyn. On the contrary, coarse-grained computations and a combination of short all-atom simulations predicted a rigid α C-helix leading to a two-step activation mechanism requiring the loop opening to facilitate the helix rotation to the inward conformation.^{17,32,68,69} As a consequence, a closed loop with an inward pointing Glu310 residue has never been reported for c-Src, while it is consistently present in both our C and C* free energy basins. Our model shows that the combination of an

α C-helix and Glu310 pointing inward with a folded A-loop is accessible and realistic from a steric point of view. However, without further experimental validation, we cannot rule out that this peculiar combination is due to a slight unbalance of the interactions between the two conformations in the model used.

4. CONCLUSIONS

We describe a novel hybrid all-atom structure-based model able to reproduce the folding transition of both helical and β -sheet globular proteins in good agreement with the available experimental and computational data. The inclusion of multiple reference structures makes it useful in characterizing large-scale conformational transitions as the activation transition of the A-loop of the catalytic domain of c-Src. The major approximation of structure-based models, namely, the absence of any non-native favorable interactions, has been partially addressed in our model by including transferable local interactions that are independent of the reference structures. Albeit the model is limited by the need of an a priori knowledge of one or more reference structures, it can still be applied to study a plethora of biological phenomena of interest for which the crystal structures are available. The approach is much faster than all-atom MD with an explicit solvent; still it has a predictive value that goes well beyond the reference structures that are used to define it. For instance, an accurate reproduction of the transition states of the villin and SH3 folding are indicative of the correct description of a complex and realistic conformational ensemble of metastable structures. Moreover, the emergence of a third minimum in the free energy landscape of the conformational transition of c-Src, not included a priori in the model but in accordance with the proposed stepwise activation mechanism, further shows the versatility of the model. Altogether, these results suggest that the proposed hybrid model can be a viable approach to studying large-scale conformational transitions in proteins.

APPENDIX

To determine the value of the ϵ_0 interaction parameter, we first estimated its scale knowing that the van der Waals interaction energy between two carbon atoms in the AMBER force field is about 0.6 kJ/mol (this value turned out to be very small for our model). We then proceeded to the identification of a reasonable range of values through the visual inspection of short molecular dynamics trajectories at different values of ϵ_0 . Mainly, two behaviors have been observed: either a fast unfolding took place resulting from an ϵ_0 value too small or, on the contrary, the atoms' fluctuations were very small around the starting native conformation, which suggested a too large value of ϵ_0 . Once a reasonable range has been identified (3.0–4.0), we proceeded to a more quantitative analysis.

For each of the values of ϵ_0 (3.0, 3.2, 3.5, 3.8, 4.0), we simulated the SH3 protein using the parallel tempering scheme to sample a large range of temperatures. We compared both the temperature corresponding to the peak in the specific heat to the experimental folding temperature and the correlation of the atom fluctuations with a classical, explicit solvent, AMBER99SB-ILDN force field³⁷ simulation of the same system at $T = 350$ K.

As a result, even though a value of $\epsilon_0 = 3.5$ would have given a lower folding temperature for SH3, closer to the experimental one $T_f^{\text{exp}} \approx 350$ K,²² we eventually chose $\epsilon_0 = 3.8$ because it better reproduced the atom fluctuations as well as resulted in a faster folding protein (see Table 2). We therefore think that in the small

Table 2. Comparison of Different Value of the Interaction Parameter ϵ_0 .^a

ϵ_0	T_f	$r_{C\alpha-RMSF}$
3.0	335 K	0.78
3.2	336 K	0.65
3.5	348 K	0.82
3.8	357 K	0.84
4.0	n/a	0.84

^aFor each value of ϵ_0 are indicated the calculated SH3 folding temperature T_f as the peak temperature of the C_v and the Pearson's correlation coefficient $r_{C\alpha-RMSF}$ of the C_α root mean square fluctuation (RMSF) between a short standard molecular dynamics simulation in explicit solvent and our model. For the case of $\epsilon_0 = 4.0$, our simulation time did not allow for a sufficient sampling of the folding/unfolding transition.

range (3.5–3.8) the model would give quantitatively similar, and qualitatively identical, results.

AUTHOR INFORMATION

Corresponding Author

*E-mail: flgervasio@cnio.es.

ACKNOWLEDGMENT

We acknowledge support by the Spanish Ministry of Science and Innovation (MICINN) grant (BIO2010-20166, "AlteredDynamics") and the Barcelona Supercomputing Center for computer time. LS acknowledges support by a Juan de la Cierva Research Fellowship from MICINN. IM acknowledges support by a La Caixa Fellowship. The authors acknowledge G. Tiana for the implementation of the multiple histogram method.

REFERENCES

- Tozzini, V. *Curr. Opin. Struct. Biol.* **2005**, *15*, 144–150.
- Clementi, C. *Curr. Opin. Struct. Biol.* **2008**, *18*, 10–15.
- Hills, R. D.; Lu, L.; Voth, G. A. *PLoS Comput. Biol.* **2010**, *6*, e1000827.
- Monticelli, L.; Kandasamy, S. K.; Periole, X.; Larson, R. G.; Tieleman, D. P.; Marrink, S.-J. *J. Chem. Theory Comput.* **2008**, *4*, 819–834.
- Laio, A.; Gervasio, F. *Rep. Prog. Phys.* **2008**, *71*, 126601.
- Chipot, C.; Pohorille, A. *Free Energy Calculations*; Springer: New York, 2007.
- Ueda, Y.; Taketomi, H.; Gō, N. *Biopolymers* **1978**, *17*, 1531–1548.
- Whitford, P. C.; Noel, J. K.; Gosavi, S.; Schug, A.; Sanbonmatsu, K. Y.; Onuchic, J. N. *Proteins* **2009**, *75*, 430–441.
- Sutto, L.; Tiana, G.; Broglia, R. A. *Protein Sci.* **2006**, *15*, 1638–1652.
- Derreumaux, P.; Mousseau, N. *J. Chem. Phys.* **2007**, *126*, 025101.
- Hills, R. D.; Brooks, C. L. *IJMS* **2009**, *10*, 889–905.
- Bryngelson, J. D.; Onuchic, J. N.; Socci, N. D.; Wolynes, P. G. *Proteins* **1995**, *21*, 167–195.
- Camilloni, C.; Sutto, L. *J. Chem. Phys.* **2009**, *131*, 245105.
- Best, R. B.; Chen, Y.-G.; Hummer, G. *Structure* **2005**, *13*, 1755–1763.
- Okazaki, K.-i.; Koga, N.; Takada, S.; Onuchic, J. N.; Wolynes, P. G. *Proc. Natl Acad. Sci. U.S.A.* **2006**, *103*, 11844–11849.
- Zhang, B. W.; Jasnow, D.; Zuckerman, D. M. *Proc. Natl Acad. Sci. U.S.A.* **2007**, *104*, 18043–18048.
- Yang, S.; Roux, B. *PLoS Comput. Biol.* **2008**, *4*, e1000047.
- Li, W.; Wolynes, P. G.; Takada, S. *Proc. Natl Acad. Sci. U.S.A.* **2011**, *108*, 3504–3509.
- de Marco, G.; Várnai, P. *Phys. Chem. Chem. Phys.* **2009**, *11*, 10694.
- Pogorelov, T. V.; Luthey-Schulten, Z. *Biophys. J.* **2004**, *87*, 207–214.
- Meinke, J.; Hansmann, U. J. *Phys.: Condens. Matter* **2007**, *19*, 285215.
- Grantcharova, V. P.; Baker, D. *Biochemistry* **1997**, *36*, 15685–15692.
- Riddle, D. S.; Grantcharova, V. P.; Santiago, J. V.; Alm, E. J.; Ruczinski, I.; Baker, D. *Nat. Struct. Biol.* **1999**, *6*, 1016–1024.
- Grantcharova, V. P.; Riddle, D. S.; Baker, D. *Proc. Natl Acad. Sci. U.S.A.* **2000**, *97*, 7084–7089.
- Beauchamp, K. A.; Ensign, D. L.; Das, R.; Pande, V. S. *Proc. Natl Acad. Sci. U.S.A.* **2011**, *108*, 12734–12739.
- Reiner, A.; Henklein, P.; Kiefhaber, T. *Proc. Natl Acad. Sci. U.S.A.* **2010**, *107*, 4955–4960.
- Hu, K.-N.; Yau, W.-M.; Tycko, R. *J. Am. Chem. Soc.* **2010**, *132*, 24–25.
- Havlin, R.; Tycko, R. *Proc. Natl Acad. Sci. U.S.A.* **2005**, *102*, 3284.
- Kubelka, J.; Eaton, W. A.; Hofrichter, J. *J. Mol. Biol.* **2003**, *329*, 625–630.
- Ishizawar, R.; Parsons, S. *J. Cancer Cell* **2004**, *6*, 209–214.
- Gan, W.; Yang, S.; Roux, B. *Biophys. J.* **2009**, *97*, L8–L10.
- Yang, S.; Banavali, N. K.; Roux, B. *Proc. Natl Acad. Sci. U.S.A.* **2009**, *106*, 3776–3781.
- Young, M.; Gonfloni, S.; Superti-Furga, G.; Roux, B.; Kuriyan, J. *Cell* **2001**, *105*, 115–126.
- Xu, W.; Doshi, A.; Lei, M.; Eck, M.; Harrison, S. *Mol. Cell* **1999**, *3*, 629–638.
- Gonfloni, S.; Williams, J. C.; Hattula, K.; Weijland, A.; Wierenga, R. K.; Superti-Furga, G. *EMBO J.* **1997**, *16*, 7261–7271.
- Cowan-Jacob, S. W.; Fendrich, G.; Manley, P. W.; Jahnke, W.; Fabbro, D.; Liebetanz, J.; Meyer, T. *Structure* **2005**, *13*, 861–871.
- Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; Shaw, D. E. *Proteins* **2010**, *78*, 1950–1958.
- Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. *J. Am. Chem. Soc.* **1995**, *117*, 5179–5197.
- Wu, L.; Zhang, J.; Qin, M.; Liu, F.; Wang, W. *J. Chem. Phys.* **2008**, *128*, 235103.
- Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. *J. Chem. Theory Comput.* **2008**, *4*, 435–447.
- Hansmann, U. H. E. *Chem. Phys. Lett.* **1997**, *281*, 140–150.
- Piana, S.; Laio, A. *J Phys Chem B* **2007**, *111*, 4553–4559.
- Bonomi, M.; Branduardi, D.; Bussi, G.; Camilloni, C.; Provasi, D.; Raiteri, P.; Donadio, D.; Marinelli, F.; Pietrucci, F.; Broglia, R. *Comput. Phys. Commun.* **2009**, *180*, 1961–1972.
- Shaw, D. E.; Maragakis, P.; Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Eastwood, M. P.; Bank, J. A.; Jumper, J. M.; Salmon, J. K.; Shan, Y.; Wrighers, W. *Science* **2010**, *330*, 341–346.
- Berteotti, A.; Cavalli, A.; Branduardi, D.; Gervasio, F. L.; Recanatini, M.; Parrinello, M. *J. Am. Chem. Soc.* **2009**, *131*, 244–250.
- Ferrenberg, A.; Swendsen, R. *Phys. Rev. Lett.* **1989**, *63*, 1195–1198.
- Ferkinghoff-Borg, J. *Eur. Phys. J. B* **2002**, *29*, 481–484.
- Kaya, H.; Chan, H. S. *Proteins* **2000**, *40*, 637–661.
- Lazaridis, T.; Karplus, M. *Science* **1997**, *278*, 1928–1931.
- Li, A.; Daggett, V. *Proc. Natl Acad. Sci. U.S.A.* **1994**, *91*, 10430–10434.
- McKnight, C. J.; Doering, D. S.; Matsudaira, P. T.; Kim, P. S. *J. Mol. Biol.* **1996**, *260*, 126–134.
- Freddolino, P. L.; Schulten, K. *Biophys. J.* **2009**, *97*, 2338–2347.
- Yang, J. S.; Wallin, S.; Shakhnovich, E. I. *Proc. Natl Acad. Sci. U.S.A.* **2008**, *105*, 895–900.
- Lei, H.; Duan, Y. *J. Mol. Biol.* **2007**, *370*, 196–206.
- Lei, H.; Wu, C.; Liu, H.; Duan, Y. *Proc. Natl Acad. Sci. U.S.A.* **2007**, *104*, 4925.
- Bunagan, M. R.; Gao, J.; Kelly, J. W.; Gai, F. *J. Am. Chem. Soc.* **2009**, *131*, 7470–7476.

- (57) Vermeulen, W.; van Troys, M.; Bourry, D.; Dewitte, D.; Rossenu, S.; Goethals, M.; Borremans, F. A. M.; Vandekerckhove, J.; Martins, J. C.; Ampe, C. *J. Mol. Biol.* **2006**, *359*, 1277–1292.
- (58) Lam, A. R.; Borreguero, J. M.; Ding, F.; Dokholyan, N. V.; Buldyrev, S. V.; Stanley, H. E.; Shakhnovich, E. *J. Mol. Biol.* **2007**, *373*, 1348–1360.
- (59) Mitomo, D.; Nakamura, H. K.; Ikeda, K.; Yamagishi, A.; Higo, J. *Proteins* **2006**, *64*, 883–894.
- (60) Parsons, S. J.; Parsons, J. T. *Oncogene* **2004**, *23*, 7906–7909.
- (61) Xu, W.; Harrison, S.; Eck, M. *Nature* **1997**, *385*, 595–602.
- (62) Rosen, N.; Bolen, J. B.; Schwartz, A. M.; Cohen, P.; DeSeau, V.; Israel, M. A. *J. Biol. Chem.* **1986**, *261*, 13754–13759.
- (63) Verbeek, B. S.; Vroom, T. M.; Adriaansen-Slot, S. S.; Ottenhoff-Kalff, A. E.; Geertzema, J. G.; Hennipman, A.; Rijksen, G. *J. Pathol.* **1996**, *180*, 383–388.
- (64) Bose, R.; Holbert, M. A.; Pickin, K. A.; Cole, P. A. *Curr. Opin. Struct. Biol.* **2006**, *16*, 668–675.
- (65) Porter, M.; Schindler, T.; Kuriyan, J.; Miller, W. T. *J. Biol. Chem.* **2000**, *275*, 2721–2726.
- (66) Adams, J. A. *Biochemistry* **2003**, *42*, 601–607.
- (67) Ozkirimli, E.; Post, C. B. *Protein Sci.* **2006**, *15*, 1051–1062.
- (68) Dixit, A.; Verkhivker, G. M. *PLoS Comput Biol* **2009**, *5*, e1000487.
- (69) Banavali, N. K.; Roux, B. *Proteins* **2007**, *67*, 1096–1112.

Detailed Investigation of the OH Radical Quenching by Natural Antioxidant Caffeic Acid Studied by Quantum Mechanical Models

Monica Leopoldini, Sandro G. Chiodo, Nino Russo,* and Marirosa Toscano

Dipartimento di Chimica and Centro di Calcolo ad Alte Prestazioni per Elaborazioni Parallele e Distribuite-Centro d'Eccellenza MIUR, Università della Calabria, I-87030 Arcavacata di Rende (CS), Italy

S Supporting Information

ABSTRACT: The effectiveness of naturally occurring antioxidant caffeic acid in the inactivation of the very damaging hydroxyl radical has been theoretically investigated by means of hybrid density functional theory. Three possible pathways by which caffeic acid may inactivate free radicals were analyzed: hydrogen abstraction from all available hydrogen atoms, hydroxyl radical addition to all carbon atoms in the molecule, and single electron transfer. The reaction paths were traced independently, and the respective thermal rate constants were calculated using variational transition-state theory including the contribution of tunneling. The more reactive sites in caffeic acid are the C₄OH phenolic group and the C₄ carbon atom, for the hydrogen abstraction and radical addition, respectively. The single electron transfer process seems to be thermodynamically unfavored, in both polar and nonpolar media. Both hydrogen abstraction and radical addition are very feasible, with a slight preference for the latter, with a rate constant of $7.29 \times 10^{10} \text{ M}^{-1} \text{ s}^{-1}$ at 300 K. Tunnel effects are found to be quite unimportant in both cases. Results indicate caffeic acid as a potent natural antioxidant in trapping and scavenging hydroxyl radicals.

1. INTRODUCTION

Phenolic compounds are plant secondary metabolites, and they are commonly found in herbs and fruits such as berries, apples, citrus fruit, cocoa, and grapes; vegetables like onions, olives, tomatoes, broccoli, lettuce, and soybeans; grains and cereals; green and black teas; coffee beans; propolis; and red and white wines.^{1–8}

Phenolic compounds are synthesized in the secondary metabolism of plants from two major synthetic pathways: the shikimate and the acetate pathway.⁹

Plants have evolved to produce phenolic compounds compounds to protect against fungal parasites,¹⁰ herbivores, pathogens, and oxidative cell injury.¹¹ Furthermore, they produce stimuli to assist in pollination¹² and guide insects to their food source.¹⁰

In recent decades, phenolic compounds have attracted growing global interest upon the discovery of the so-called “French Paradox”, i.e., the observation that although the French have a tendency to smoke and a diet rich in fats, they show much reduced rates of coronary heart diseases when compared with other northern European nations.¹³ The most popular explanation has been recognized in the relatively high daily consumption of red wines rich in phenolic compounds by the French, which in some way protects them from heart disease.^{1,14}

The term phenolics encompasses approximately 8000 naturally occurring compounds, all having a phenol ring. A further classification divides them in polyphenols and simple phenols, depending on the number of phenol subunits. Simple phenols include phenolic acids.¹⁵ Polyphenols possessing at least two phenol subunits include the flavonoids and the stilbenes. Those compounds with three or more phenol subunits are referred to as the tannins.¹⁶

The pharmacological, medicinal, and biochemical properties of phenolics have been extensively reviewed.^{17–19} They have

been reported to have antioxidant,²⁰ vasodilatory, anticarcinogenic, antiinflammatory, immune-stimulating, antiallergic, and antiviral and estrogenic effects.^{21,22} In addition, inhibition activities against several enzymes like phospholipase A₂, cyclooxygenase, lipoxygenase,^{23–30} glutathione reductase,³¹ and xanthine oxidase enzymes³² were proven.

A diet rich in fruits and vegetables that contain these compounds reduces the risk of cardiovascular and some other diseases, and most countries have developed recommendations for an increased intake of fruit and vegetables.³³

The best described property of phenolics is their antioxidant capability toward free radicals normally produced in cell metabolism or in response to external factors. Free radicals can damage biomolecules such as lipids, nucleic acids, and proteins; cause cellular membranes peroxidation;^{34,35} and attract various inflammatory mediators.³⁶

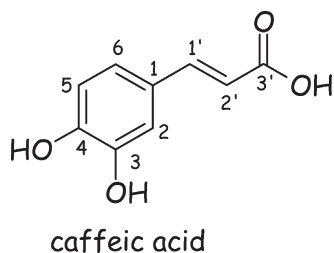
Phenolic acids are phenols characterized by a carboxylic functionality. They contain two distinguishing constitutive carbon frameworks: the hydroxycinnamic and hydroxybenzoic structures. Hydroxycinnamic acids are more common than hydroxybenzoic ones and consist mainly of *p*-coumaric, sinapic, ferulic, and caffeic acids;¹⁵ the latter is thought to be the most abundant in the diet.³

Phenolic acids exist primarily as conjugates of sugars, polysaccharides, or organic acids, whereas the free forms are less frequently observed in nature. The quantitatively most important conjugate of caffeic acid is its ester with quinic acid, 5-caffeoyl-quinic acid (also known as chlorogenic acid). High concentrations of phenolic acids are found in coffee, apples, citrus fruits and juices, and the bran of cereal grains.³ Caffeic acid seems to

Received: August 16, 2011

Published: November 08, 2011

Scheme 1. Structure of Caffeic Acid



contribute to color stability and protection against oxidation in the red wines.³⁷

Recent experimental evidence suggests caffeic acid to have a beneficial impact *in vivo*.^{38–40} Caffeic acid has biological and pharmacological properties, such as antiviral,⁴¹ antioxidant,⁴² anti-inflammatory,⁴³ anticarcinogenic,⁴⁴ and immunomodulatory activities.⁴⁵ In addition, it completely blocks the production of reactive oxygen species (ROS).⁴⁶

According to a recent research,⁴⁷ caffeic acid was a superior antioxidant compared with *p*-coumaric and ferulic acids in inhibiting LDL oxidation^{48,49} but also in quenching radicals⁵⁰ and singlet oxygen.

The molecular basis for the antioxidant properties of phenolic compounds is recognized in three main mechanisms, arising from the direct reaction with free radicals.^{51–54} Another indirect modus operandi comes from their ability to chelate free metals that are involved in reactions finally generating free radicals.^{55,56}

As primary antioxidants, polyphenols inactivate free radicals according to the hydrogen atom transfer (HAT; eq 1), to the radical adduct formation (RAF; eq 2), and to the single electron transfer (SET; eq 3) mechanisms. In mechanism 1, the antioxidant, ArOH, reacts with the free radical, R[•], by transferring to it a hydrogen atom, through homolytic rupture of the O–H bond:



The radical adduct formation (RAF) mechanism (2) provides for the formation of an adduct between the radical and the antioxidant:



The SET mechanism (3) provides for an electron to be donated to the R[•]:



The products of mechanisms 1, 2, and 3 (ArO[•], ArOH–R[•], and ArOH^{+•}) are aromatic structures in which the odd electron, originated through the reaction with the free radical, has the possibility to be spread over the entire molecule, resulting in a significant radical stabilization.^{51–54}

On the basis of the reaction rate constants of polyphenols with several radicals and the stability of the polyphenol-derived radicals, the antioxidant capability of some of these compounds was evaluated. Among the hydroxycinnamates, caffeic acid (see scheme 1) presented high rate constants toward several types of oxidant species like ROO[•] ($1.5 \times 10^7 \text{ M}^{-1} \text{ s}^{-1}$),⁵⁷ O₂^{•–} ($0.96 \times 10^6 \text{ M}^{-1} \text{ s}^{-1}$), HO[•] ($3.24 \times 10^9 \text{ M}^{-1} \text{ s}^{-1}$),⁵⁸ and ¹O₂ ($5.1 \times 10^6 \text{ M}^{-1} \text{ s}^{-1}$).⁵⁹

The caffeic acid oligomers have been shown to be very effective scavengers of 2,2-diphenyl-1-picryl hydrazyl (DPPH) and superoxide anion radicals.^{60,61}

On the basis of what was previously mentioned, it is evident that further research studies are needed and would be useful to fully understand the antioxidant capability of phenolics and to individuate a feasible molecular mechanism through which these plant metabolites scavenge free radicals.

The aim of this work is to study the reactivity of caffeic acid toward the OH radical, accepted to be one of the most reactive among the ROSs, with a very short half-life (around 10^{-9} s) and high reactivity. Different mechanisms and all possible sites of reaction/attack have been examined. Thermodynamic and kinetic data have been collected to identify the main channels of reaction.

2. METHODS

Geometry optimization and frequency calculations were performed by means of density functional theory (DFT). The Becke exchange and Lee, Yang, and Parr correlation (B3LYP)^{62–65} and the M05-2X⁶⁶ functionals were chosen. The 6-31+G(d) basis set^{67–70} was used for the representation of the C, O, and H atoms.

The M05-2X functional has been recommended for applications in thermochemistry, kinetics, and noncovalent interactions by its developers,⁶⁶ and it has been also successfully used by independent authors.^{71–77} However, B3LYP still represents the most widely used DFT functional, so it is used in this study for purpose of comparison.

Vibrational frequencies were obtained at the M05-2X and B3LYP levels, in order to compute the zero point energy corrections and to determine the nature of all stationary points, as minima and saddle points.

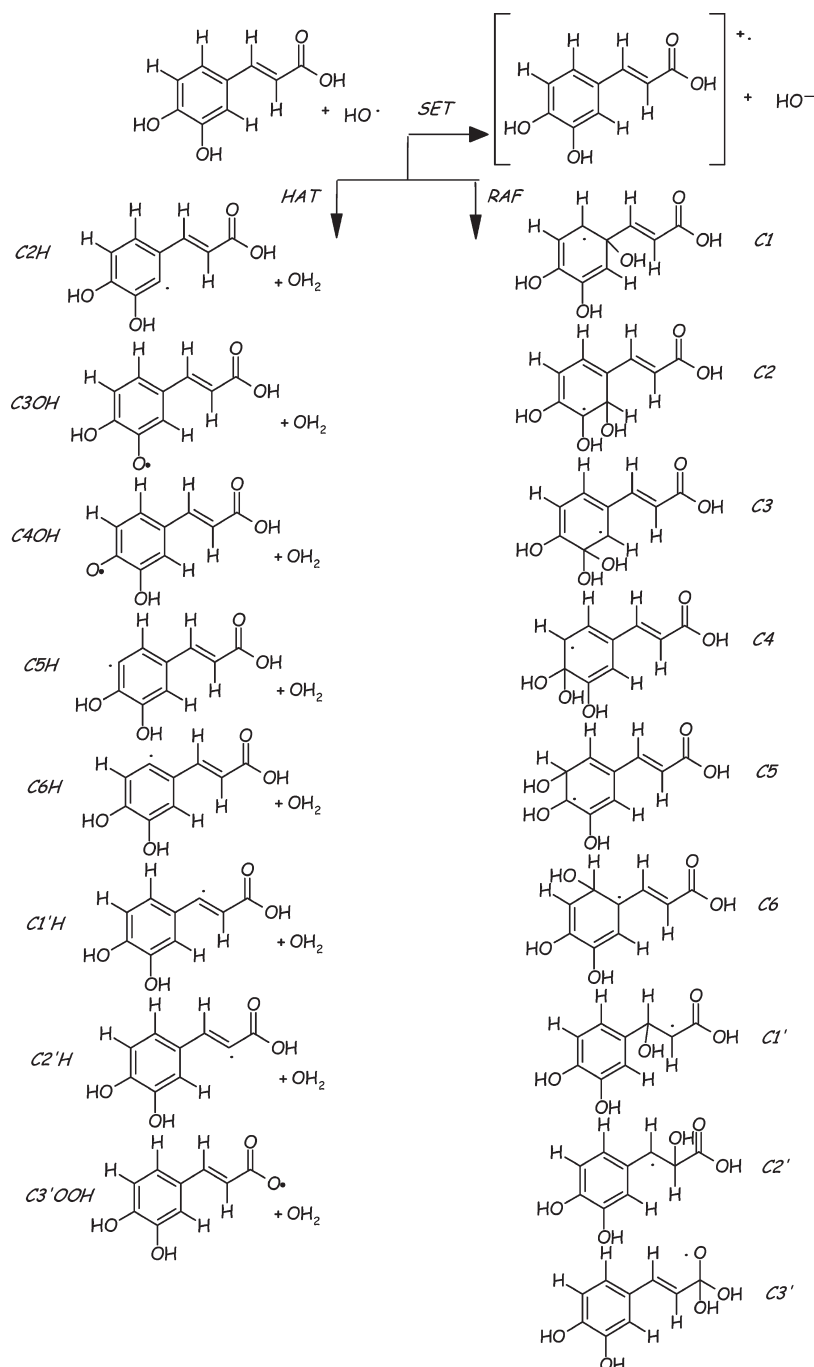
To confirm that a given transition state connects reactants and products and to carry out the dynamical computations, the intrinsic reaction coordinate (IRC) or minimum energy path (MEP)^{78,79} was constructed at the M05-2X level, starting from the respective saddle point geometry and going downhill to both the asymptotic reactant and the product channels, in an iso-inertial mass-weighted Cartesian coordinate system. A step size, δs , of 0.02 bohr (where *s* is the distance along the MEP which starts from 0 for the saddle point and assumes negative and positive values for reactants and products, respectively) was used in all cases. The second derivative matrix was computed at each of three points along the reaction path.

Solvent effects were introduced as single point computations on the optimized gas phase structures in the framework of the self-consistent reaction field polarizable continuum model (SCRF-C-PCM)^{80–82} in which the cavity is created *via* a series of overlapping spheres. The united atom topological model (UA0) applied on atomic radii of the UFF force field⁸³ was used to build the cavity, in the gas-phase equilibrium geometry. The dielectric constant values $\epsilon = 78.3553$ and $\epsilon = 2.2706$ were chosen to reproduce water and benzene media, respectively.

All rate constants were estimated by using canonical variational transition state theory (CVT),^{84–88} corrected by the semiclassical multidimensional small-curvature tunnelling (SCT) approach.^{89–92} The CVT/SCT method has been chosen since it has been successfully applied on the reactions of radical trapping performed by natural antioxidants.^{93–96}

All of the calculations were performed using the Gaussian 03⁹⁷ and Polyrate9.7⁹⁸ codes.

Scheme 2. Possible Reaction Mechanisms



3. RESULTS AND DISCUSSION

3.1. OH Scavenging by Caffeic Acid in the Neutral Form: Geometries and Energies. In Scheme 2, a representation of all possible reaction channels and attack sites for the trapping of the $\cdot\text{OH}$ by caffeic acid is shown.

Hydrogen atom transfer (HAT) can occur from the C_2H , C_3OH , C_4OH , C_5H , C_6H , $\text{C}_1'\text{H}$, $\text{C}_2'\text{H}$, and $\text{C}_3'\text{OOH}$ sites and radical adduct formation (RAF) at the C_1 , C_2 , C_3 , C_4 , C_5 , C_6 , C_1' , C_2' , and C_3' positions. In single electron transfer (SET), the one electron is transferred from the HOMO of the phenolic molecule to the hydroxyl radical.

In Tables 1 and 2, the relative energies of the stationary points (reactant complex (RC), transition state (TS), product complex (PC), and products (P) for the HAT and reactant complex (RC), transition state (TS), and product (P) for the RAF) with respect to the reactants (R) are reported for all mechanisms investigated, considering all positions of reactivity. In Figures 1, 2, 3, 4 and 5, the M05-2X and B3LYP optimized geometries for stationary points corresponding to the HAT, RAF, and SET mechanisms are shown.

Caffeic acid is a planar species, at both the M05-2X and B3LYP levels, characterized by electronic conjugation and delocalization that concern the whole molecule up to the carboxylic moiety.

Table 1. M052X/6-31+G(d) and B3LYP/6-31+G(d) Relative Energies (in kcal/mol) of the Stationary Points for the HAT Mechanism, Considering Both the Neutral and the Deprotonated (In Parentheses) Forms of Caffeic Acid

site	HAT M052X			
	RC	TS	PC	P
C ₂ H	-6.57 (2.13)	7.56 (19.0)	-4.31 (3.38)	1.63 (11.93)
C ₃ H	-7.32 (/)	4.38 (/)	-10.56 (/)	0.32 (/)
C ₆ H	-4.66 (-9.96)	8.70 (19.0)	-3.14 (-4.54)	1.01 (11.93)
C ₁ 'H	-8.92 (-9.92)	3.62 (2.89)	-14.12 (-17.32)	-5.70 (0.24)
C ₂ 'H	-5.85 (-9.92)	7.19 (3.51)	-5.98 (-13.09)	-0.20 (0.24)
C ₃ OH	-5.96 (/)	3.03 (/)	-47.03 (-43.93)	-35.42 (-32.52)
C ₄ OH	-5.79 (/)	-1.00 (/)	-45.89 (-49.01)	-37.84 (-39.08)
C ₃ 'OOH	-10.86(/)	0.24 (/)	2.45 (/)	-2.50 (/)

site	HAT B3LYP			
	RC	TS	PC	P
C ₂ H	-5.53 (-8.93)	5.52 (3.84)	3.77 (-7.69)	0.62 (0.02)
C ₃ H	-1.72 (/)	1.41 (/)	-10.13 (/)	-0.66 (/)
C ₆ H	-4.96 (-19.72)	5.80 (0.75)	-9.88 (-15.75)	6.27 (-0.55)
C ₁ 'H	-8.09 (-19.67)	0.85 (-10.43)	-14.21 (-28.39)	-7.81 (-13.04)
C ₂ 'H	-5.14 (-20.11)	4.00 (-11.36)	-6.13 (-24.67)	-2.16 (-9.93)
C ₃ OH	/ (/)	/ (/)	-49.79 (-59.99)	-39.35 (-49.17)
C ₄ OH	/ (/)	/ (/)	-48.78 (-66.29)	-42.09 (-56.21)
C ₃ 'OOH	-9.92 (/)	-5.39 (/)	-14.92 (/)	-10.47 (/)

Table 2. M052X/6-31+G(d) and B3LYP/6-31+G(d) Relative Energies (in kcal/mol) of the Stationary Points for the RAF Mechanism, Considering Both the Neutral and the Deprotonated (In Parentheses) Forms of Caffeic Acid

site	RAF M052X		
	RC	TS	P
C ₁	-1.06 (-9.96)	1.64 (12.30)	-16.20 (-8.31)
C ₂	-4.48 (-9.91)	-2.35 (5.09)	-26.64 (-18.35)
C ₃	-4.42 (-1.51)	-1.59 (5.58)	-24.76 (-15.90)
C ₄	-7.35 (/)	-5.99 (/)	-34.49 (-27.62)
C ₅	-7.36 (-9.84)	-4.24 (0.54)	-24.16 (-16.80)
C ₆	-4.69 (-9.83)	-2.41 (4.04)	-24.42 (-17.91)
C ₁ '	-4.80 (-9.96)	-3.41 (-2.87)	-35.41 (-35.05)
C ₂ '	-8.92 (-9.84)	-3.29 (-3.10)	-36.48 (-43.00)
C ₃ '	-10.88 (-10.50)	9.33 (18.38)	-4.63 (3.01)

site	RAF B3LYP		
	RC	TS	P
C ₁	-5.13 (-19.72)	0.18 (-0.74)	-10.52 (-13.78)
C ₂	-4.28 (-19.70)	-4.27 (-8.48)	-22.77 (-25.36)
C ₃	/ (-13.28)	/ (-8.43)	-19.35 (-21.25)
C ₄	/ (/)	/ (/)	-29.06 (-33.40)
C ₅	-6.63 (-19.71)	-5.98 (-14.39)	-20.37 (-23.83)
C ₆	-3.94 (-19.70)	-3.04 (-10.41)	-21.25 (-25.16)
C ₁ '	-8.04 (-19.72)	-4.92 (-15.07)	-28.38 (-39.53)
C ₂ '	-8.07 (-19.51)	-2.29 (-15.47)	-30.27 (-48.08)
C ₃ '	-9.92 (-20.62)	8.04 (6.23)	-4.33 (-8.04)

In the minimum energy structure, the 3OH and 4OH groups are arranged in such a way as to realize a hydrogen bond.

All HAT reactant complex (RC) stationary points are characterized by a significant hydrogen bond established between the $\cdot\text{OH}$ and the hydrogen present on the site where the abstraction should occur.

In the reactant complex RC arising from the reaction at the C₂H site, two H bonds are established between the hydroxyl radical and the C₂H and C₃OH moieties. $\cdot\text{OH}$ acts as donor in the case of the interaction with the C₃OH group (1.93 and 1.90 Å, at M05-2X and B3LYP level) and as an acceptor toward C₂H (2.67 and 3.10 Å, at the M05-2X and B3LYP levels). The reaction passes through a transition state (imaginary frequency at 1462 and 1549 cm⁻¹, considering the M05-2X and the B3LYP functionals) in which the hydrogen atom attached to the C₂ is being transferred to the hydroxyl radical (C₂H \cdots OH and C₂ \cdots HOH critical distances are 1.22 and 1.26 (M05-2X) and 1.18 and 1.31 Å (B3LYP)). M05-2X and B3LYP product complexes (PC) show the water molecule hydrogen-bonded to the C₃OH group at 2.12 and 1.97 Å, respectively.

As far as HAT at the C₃OH site is concerned, the two functionals give very different results. Namely, M05-2X finds a stationary point along the potential energy profile corresponding to a complex between reactants (see Figure 1), characterized by two hydrogen bonds involving the $\cdot\text{OH}$ and the C₃OH (1.93 Å) and the C₄OH (1.98 Å) groups. Computations employing the B3LYP functional have yielded a stationary point in which the H transfer has occurred spontaneously without passing through a saddle point and providing the product complex between the water molecule and the caffeic acid radicalized at the C₃O \cdot . This different behavior may be due to the already known bad treatment and description of nonbonded interactions by the B3LYP functional.⁹⁹ In the M05-2X transition state (3234i cm⁻¹) for the reaction at the C₃OH, the hydrogen atom is contemporaneously bonded to the C₃O and to the OH at 1.07 and 1.33 Å. In the product complex, the water molecule is hydrogen-bonded

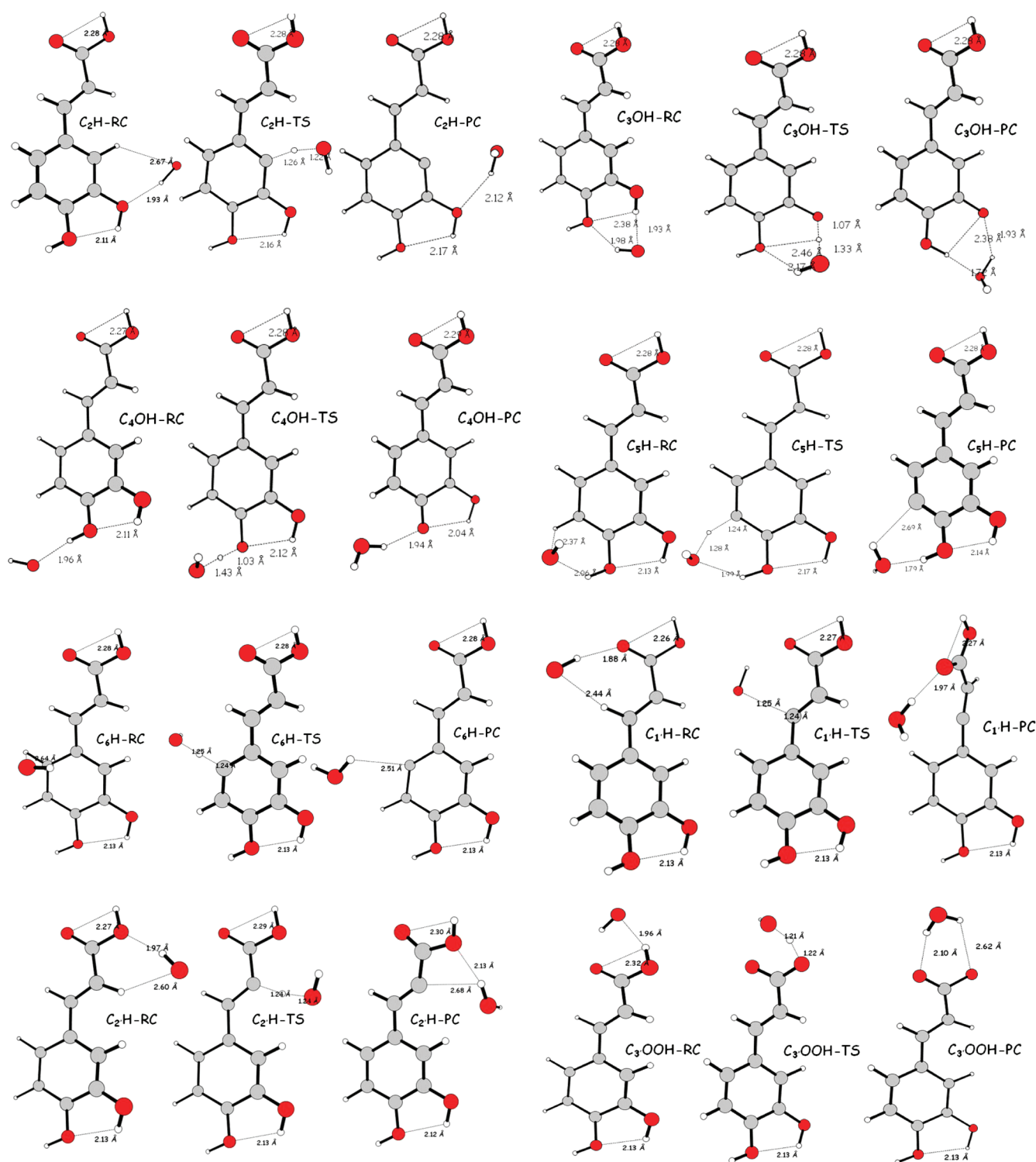


Figure 1. M05-2X optimized geometries of the stationary points encountered along the HAT mechanism.

to the radicalized C_3O at 1.93 (M05-2X) and 1.86 (B3LYP) Å. A further H-bond involves the C_3O^{\bullet} and the hydroxyl at the C_4 , which is always recognized to be the crucial feature ensuring stability to the formed radical.¹⁰⁰

Results concerning HAT at the C_4OH moiety entail similar considerations to those for the HAT at C_3OH . B3LYP fails in locating a stationary point corresponding to both the reactant complex and the transition state. The complex between reactants

is obtained at the M05-2X level when the $\bullet OH$ radical approaches the caffeic acid C_4OH group at a distance of 1.96 Å (H-bond). The H atom is then transferred to the hydroxyl radical through the M05-2X transition state TS, at the imaginary frequency of 2785 cm^{-1} corresponding to the stretching of the $C_4O \cdots H$ and $C_4OH \cdots OH$ couple of bonds (distances are 1.03 and 1.43 Å). In the product complexes at both M05-2X and B3LYP, H_2O interacts with the radicalized oxygen in the caffeic acid,

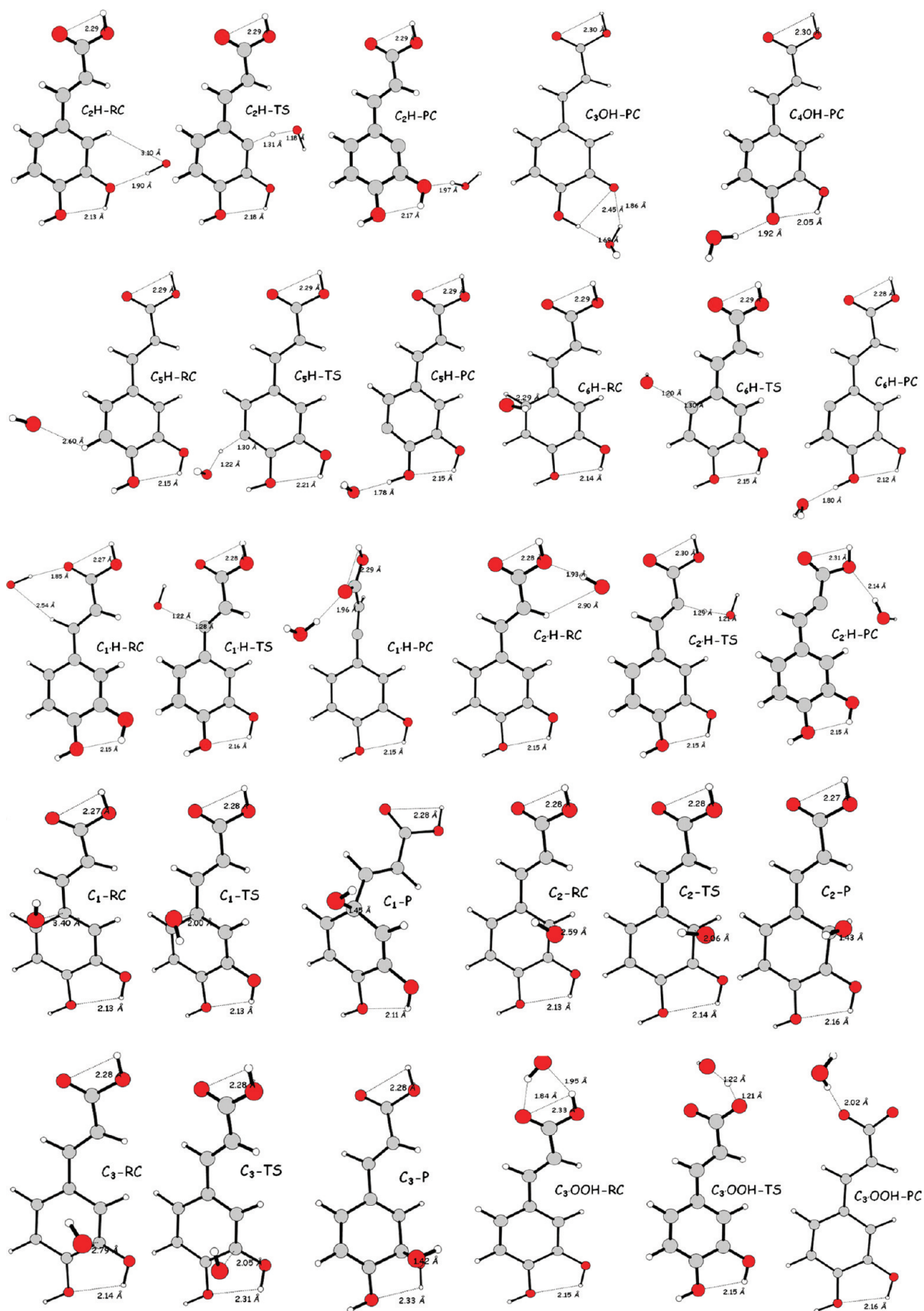


Figure 2. B3LYP optimized geometries of the stationary points encountered along the HAT mechanism.

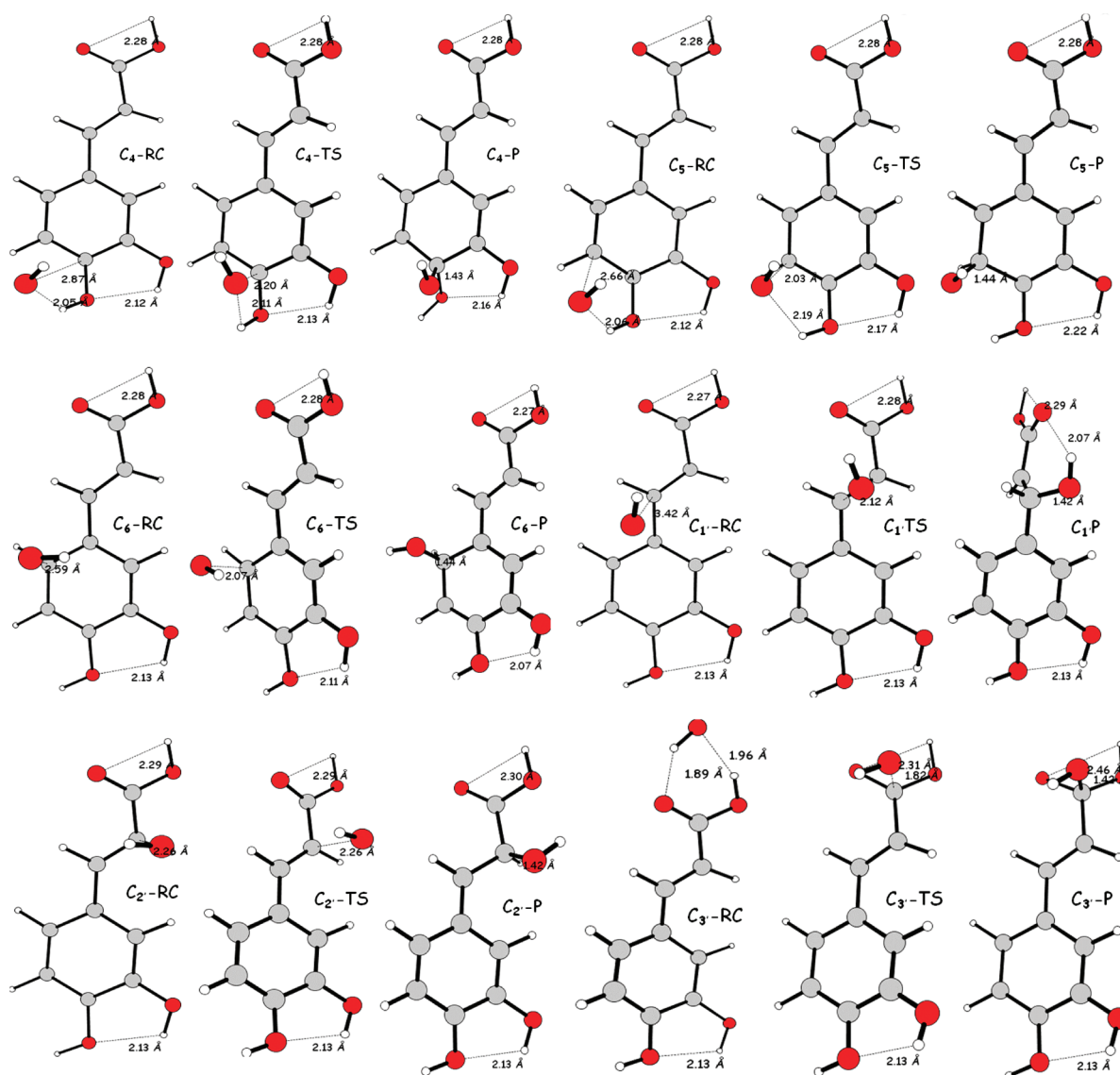


Figure 3. M05-2X optimized geometries of the stationary points encountered along the RAF mechanism.

establishing a H bond whose length is computed to be 1.94 and 1.92 Å for the two functionals, respectively. The internal hydrogen bond involving C₃OH and C₄OH is retained on going from RC to TS to PC, and it contributes to the stabilization of the electronic vacancy in the caffeic acid C₄O radical.

The hydrogen transfer process from the C₅H position starts with the formation of the reactant complex that is described differently by M05-2X and B3LYP functionals as far as the weak interactions are concerned. In the M05-2X minimized RC, [•]OH strongly interacts with the C₄OH group (2.06 Å) and to a lesser degree with the C₅H (2.37 Å). The RC within the B3LYP computations shows only a C₅H...OH hydrogen-like interaction (2.60 Å). The transition state for the H transfer is found when the C₅...HOH and C₅H...OH distances assume the values of 1.24 and 1.28 Å (M05-2X) and 1.30 and 1.22 Å (B3LYP). Its nature of the saddle point is confirmed by the imaginary frequencies at 1524 and 1580 cm⁻¹, at the M05-2X and B3LYP levels, respectively. M05-2X and B3LYP geometries for the product complex are quite similar, as shown in Figures 1 and 2.

The reactant complex between [•]OH and caffeic acid at the C₆H moiety shows a hydrogen like interaction between the hydrogen in the hydroxyl radical and the π electrons in the aromatic ring of the phenolic molecule, at both levels of theory. Passing through the transition state in which the H atom is bonded to both the carbon atom at 1.24 and 1.30 Å, and to the hydroxyl radical at 1.25 and 1.20 Å, at M05-2X and B3LYP, respectively, the reaction proceeds toward the product complex. The M05-2X PC shows the water molecule interacting with the radicalized C₆ atom (HOH...C distance is 2.51 Å), while in the B3LYP computations the water molecule forms a H bond with the C₄OH group (see Figures 1 and 2).

The hydrogen atoms present in the C=C double bond may be in principle also abstracted by the OH radical, at both the C₁' and C₂' positions.

Because of the presence of the carboxylic group, the reactant complex (RC) within the pathway at the C₁'H is characterized by a H bond with the carboxylic oxygen (1.88 and 1.85 Å, at the M05-2X and B3LYP levels, respectively). Abstraction of the C₁' bonded hydrogen atom occurs through the transition state

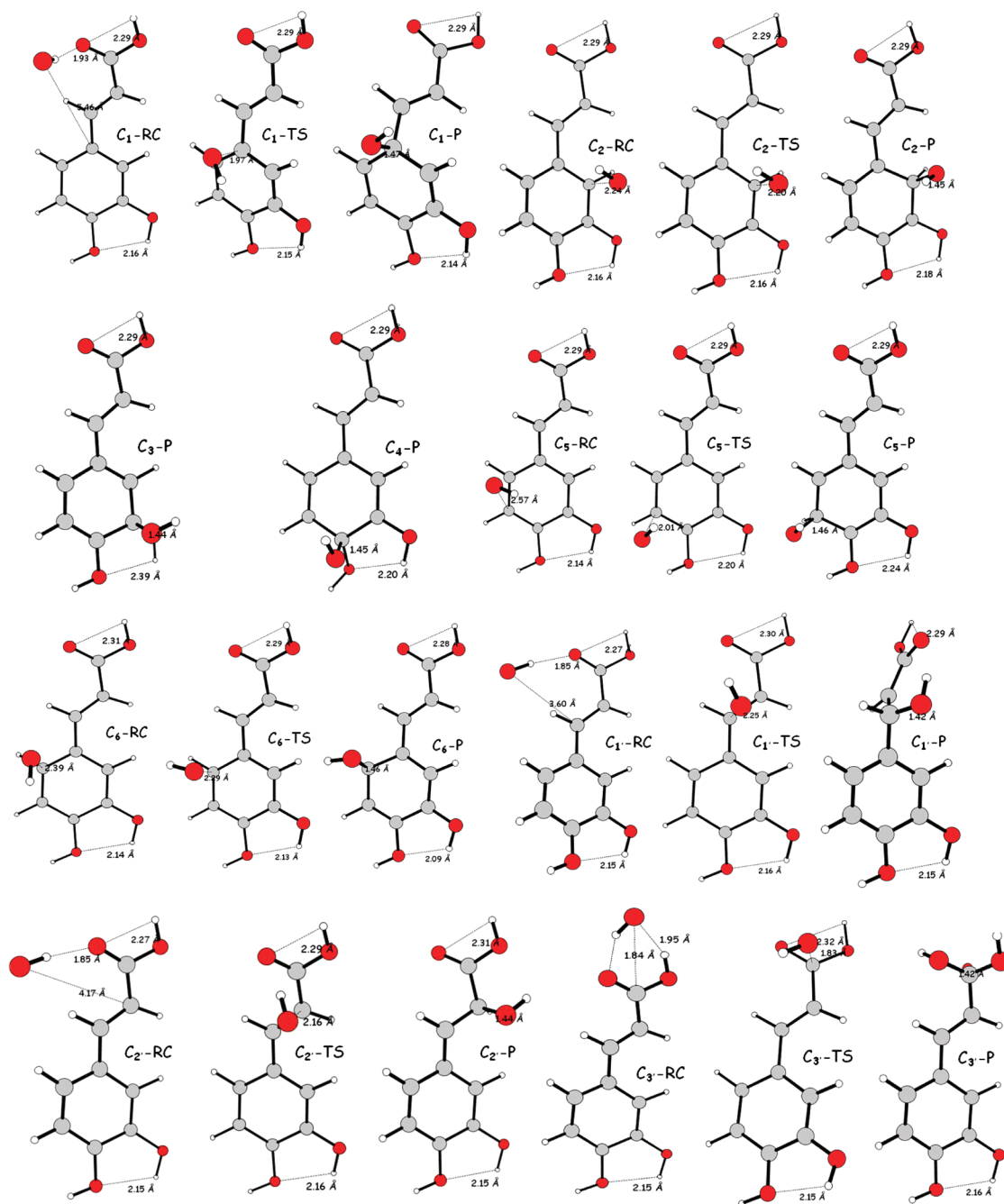


Figure 4. B3LYP optimized geometries of the stationary points encountered along the RAF mechanism.

(imaginary frequencies at 1605 and 1532 cm^{-1} , at the M05-2X and B3LYP levels, respectively) in which the critical distances $C_1 \cdots H$ and the $C_1 H \cdots OH$ assume the values of 1.24 and 1.25 Å (M05-2X) and 1.28 and 1.22 Å (B3LYP). The HAT $C_1 H$ pathway product complex shows a H bond involving the water molecule and the carboxylic oxygen in the caffeic acid (~ 1.96 Å for both functionals) and a particular rearrangement of the ethylene group perpendicular to the plane of the ring for both functionals employed (see Figures 1 and 2). This reorganization should favor resonance and conjugation effects and thus a better stabilization of the unpaired electron in the just formed radical.

The RC for the HAT $C_2 H$ reaction shows, like the HAT $C_1 H$ one, a strong interaction between the OH radical and the

carboxylic group (see Figures 1 and 2). The H atom is shared between the C_2' and the O atoms in the transition state, for which animation of the vibrational modes at 1527i and 1513i cm^{-1} frequency values proposes the stretching of the $C_2' \cdots H$ and $C_2' H \cdots OH$ bonds. M05-2X and B3LYP product complexes show the same geometrical features (see Figures 1 and 2).

As far as the H abstraction by $\cdot OH$ from the COOH moiety is concerned, the reactant complex appears to be characterized by two strong H bonds in the reaction part, as shown in Figures 1 and 2, by both employed functionals, which predict also very similar geometries for the saddle point (imaginary frequencies are 1362 and 1372 cm^{-1} , at the M05-2X and B3LYP levels). To the contrary, the structure for the product complex is quite

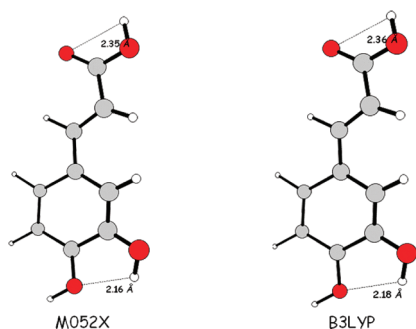


Figure 5. M05-2X and B3LYP optimized geometries of the radical cation arising from the SET mechanism.

different since the M05-2X optimized geometry shows the water molecule involved in two H bonds with the COO^\bullet group (2.10 and 2.62 Å), while B3LYP computations locate on the potential energy profile a stationary point with only one H bond (2.02 Å).

The formation of the reactant complex is predicted to be energetically favored for all reaction sites by both functionals employed (see Table 1), especially in the case of the COOH ($\Delta E = -10.86$ and -9.92 kcal/mol with respect to the unbound reactants, at M05-2X and B3LYP, respectively). M05-2X and B3LYP also give the same trend of relative energetic stability, with the only exception being the C_5H RC, which is found to lie 7.32 (M05-2X) and 1.72 (B3LYP) kcal/mol below the asymptote. This difference in the energy is due to the diverse geometrical structure for this stationary point.

Energy barriers (TS relative energies reported in Table 1) for the H transfer, determined at the M05-2X level, indicate that this process is particularly favored from the C_4OH site, followed by the reaction at C_3OOH and C_3OH . In the case of the B3LYP computations, the lack of some stationary points not located on the potential energy profile prevents us from indicating an accurate trend.

HAT processes are predicted to be thermodynamically very favored in the case of the C_3OH and C_4OH groups, as the relative energies for the PC and P stationary points suggest.

In agreement with previous works,^{52–54,96,100} results reported here indicate that the C_4OH site seems to be the most reactive one within the HAT mechanism. The better reactivity of the C_4OH group toward the $\bullet\text{OH}$ is mainly due to the stability generated by the intramolecular H bond between the two phenolic groups that is retained on going from free caffeic acid to the reactant complex, to the transition state up to the product complex and free products. This H bond helps in stabilizing the electronic deficiency generated on the phenolic oxygen during the H abstraction reaction. Additionally, in the product complex and free products coming from radicalization at the 4OH site, the unpaired electron generated is delocalized over the whole molecule thanks to the electronic conjugation characterizing the caffeic acid. Atomic spin densities for the PC of the reaction channels listed in Table S1 in the Supporting Information indicate that when OH radicalization occurs at the C_4OH position, a broad delocalization of the unpaired electron involving the C_1 , C_2 , C_3 , C_5 , C_6 , C_1' , C_2' , and C_4O atoms occurs. Radicalization at the C_3OH entails a delocalization of the odd electron only on the aromatic ring since C_1' and C_2' atoms are not involved. For the other positions, this stability cannot be achieved.

The addition of the $\bullet\text{OH}$ radical to the carbon atoms in the caffeic acid (RAF mechanism) starts again with the formation of

weak-bonded complexes in which the radical is involved in hydrogen like interactions and/or interactions with the π electrons.

The reactant complex for the RAF process at the C_1 position is characterized by an interaction of the $\bullet\text{OH}$ with the π electrons and by a H bond with the carboxylic group, as far as M05-2X and B3LYP computations are concerned. The visualization of the imaginary frequencies for the transition state (at 497 (M05-2X) and 352 (B3LYP) cm^{-1}) shows the stretching of the $\text{HO}\cdots\text{C}_1$ incoming bond. In both cases, the $-\text{CH}=\text{CH}-\text{COOH}$ chain is slightly twisted with respect to the phenyl ring (the $\text{C}_2-\text{C}_1-\text{C}_1'-\text{C}_2'$ dihedral is found to be 18.43° and 15.56° , at the M05-2X and B3LYP levels, respectively). Finally, the product is characterized by a $\text{HO}-\text{C}_1$ σ bond at 1.45 and 1.47 Å; the conjugation between the $-\text{CH}=\text{CH}-\text{COOH}$ chain and the phenyl ring has completely disappeared (see Figures 3 and 4).

Both M05-2X and B3LYP reactant complexes (RC) present a weak interaction between the aromatic ring and the $\bullet\text{OH}$ ($\text{HO}\cdots\text{C}_2$ distance is 2.59 and 2.24 Å). The transition state is located when the $\text{HO}\cdots\text{C}_2$ distance assumes the value of 2.06 (at $428i$ cm^{-1}) and 2.20 ($66i$ cm^{-1}) Å, considering the M05-2X and B3LYP computations. In the product, the HO that is now bonded to the C_2 at 1.43 (M05-2X) and 1.45 (B3LYP) Å is oriented in such a way as to realize an interaction of the hydrogen with the π electrons in the phenyl ring rather than a H bond with the hydroxyl attached at the C_3 .

Any attempt to locate the corresponding C_3 reactant complex within the RAF mechanism at the B3LYP level invariably led to structures that correspond to the final product for this channel or to the C_3OH HAT PC point. Only M05-2X locates a reactant complex on the potential energy profile, showing a $\text{HO}\cdots\text{C}_3$ distance of 2.79 Å and an interaction of the hydroxyl hydrogen with the π electrons (see Figure 3). In the M05-2X C_3 RAF transition state, the $\text{HO}\cdots\text{C}_3$ bond is forming at 2.05 Å, as indicated by the imaginary frequency at 403 cm^{-1} corresponding to the stretching of this bond. M05-2X and B3LYP product structures are shown in Figures 3 and 4 and present very similar arrangements.

Concerning the RAF at the C_4 position, again B3LYP does not find any structure corresponding to a reactant complex (as well as later to the transition state). This structure, indeed located with the M05-2X calculations for a $\text{HO}\cdots\text{C}_4$ distance of 2.87 Å, presents also a H bond with the C_4 hydroxyl. The C_4 RAF channel passes through a transition state in which the $\text{HO}\cdots\text{C}_4$ bond is forming at 2.20 Å (imaginary frequency at 247 cm^{-1}). OH finally establishes a σ bond with the C_4 carbon atom (distances are 1.43 and 1.45 Å, at M05-2X and B3LYP, respectively).

The reactant complex for the RAF at the C_5 position is very similar to the RC obtained for the RAF C_4 channel at the M05-2X level, while B3LYP computations find a stationary point in which the hydroxyl radical establishes a $\text{OH}\cdots\pi$ electron interaction (see Figures 3 and 4). The attack by the $\bullet\text{OH}$ radical on the C_5 carbon atom occurs when the $\text{HO}\cdots\text{C}_5$ distance assumes the critical value of 2.03 (M05-2X) and 2.01 (B3LYP) Å. The nature of the transition state for this saddle point is clearly indicated by the imaginary frequency at 391 and 282 cm^{-1} , at the M05-2X and B3LYP levels, respectively. The TS evolves into the product, in which the C_5-OH bond is completely formed at 1.44 and 1.46 Å.

The RC obtained for the RAF C_6 channel shows a weak interaction between the $\bullet\text{OH}$ and the C_6 carbon atom at 2.59 (M05-2X) and 2.39 (B3LYP) Å. In the transition state, the bond

$\text{OH}\cdots\text{C}_6$ measures 2.07 and 2.29 Å, at the M05-2X and B3LYP levels, respectively. The vibrational normal mode with an imaginary frequency at 435 and 207 cm^{-1} is the stretching of this bond. After the TS, a covalent product is located on the M05-2X and B3LYP potential energy profiles for a $\text{HO}-\text{C}_6$ bond of 1.44 and 1.46 Å, respectively. Only the C_6 carbon assumes the sp^3 configuration, while the other carbon atoms retain the sp^2 hybridization as it occurs for the other RAF channels (see Figures 3 and 4).

The process which leads to $\cdot\text{OH}$ radical addition to the $\text{HC}=\text{CH}$ moiety (at positions $\text{C}_{1'}$ and $\text{C}_{2'}$) or to the carboxylic group may also occur. As far as the $\text{C}_{1'}$ addition is concerned, the reactant complex (RC) presents a different arrangement depending on the exchange-correlation functional used. Particularly, B3LYP optimization leads to a RC in which the OH radical establishes a H bond with the $-\text{COOH}$ group, while the M05-2X one shows a weak interaction between the hydroxyl radical and the $-\text{CH}=\text{CH}-$ moiety. $\cdot\text{OH}$ addition occurs through the transition state TS for a $\text{HO}\cdots\text{C}_{1'}$ distance of 2.12 (M05-2X, 376i cm^{-1}) and 2.29 (B3LYP, 125i cm^{-1}) Å. The next point after the TS is the product whose geometry is predicted to be very similar by both functionals. The $\text{HO}-\text{C}_1$ bond assumes the value of 1.42 Å in both cases, with the OH pointing toward the oxygen of the $-\text{COOH}$ group (H bond is 2.07 and 1.94 Å at the M05-2X and B3LYP levels, respectively).

$\cdot\text{OH}$ addition at the $\text{C}_{2'}$ position entails the formation of the reactant complex that, at both M05-2X and B3LYP levels of theory, is characterized by a strong H bond with the carboxylic group (see Figures 3 and 4). At the saddle point for radical addition, the $\text{HO}-\text{C}_{2'}$ distance is 2.26 Å at the M05-2X level and 2.16 Å at the B3LYP level (imaginary frequencies are 283 (M05-2X) and 183 (B3LYP) cm^{-1}). The $\text{HO}-\text{C}_{2'}$ bond is completely formed in the product at 1.42 (M05-2X) and 1.44 (B3LYP) Å.

In the $\cdot\text{OH}$ addition at the $-\text{COOH}$ group, the radical approaches in the same plane of the molecule because of the formation of two H bonds with the carboxylic group (see Figures 3 and 4). The transition state sees the $\cdot\text{OH}$ attacking the carboxylic carbon atom at 1.82 (M05-2X, 663i cm^{-1}) and 1.83 (B3LYP, 497i cm^{-1}) Å and leads to the product characterized by a $\text{HO}-\text{C}_{\text{COOH}}$ bond whose length is found to be 1.42 Å by both functionals.

In Table 2, the energies for the three regions RC, TS, and P relative to the free reactants taken into account within all channels for the RAF mechanism are reported.

Reactant complex formation is energetically favored in all investigated cases, by both M05-2X and B3LYP functionals, especially considering the addition at the $\text{C}_{3'}$ position in the side chain (-10.88 and -9.92 kcal/mol, at M05-2X and B3LYP level, respectively) because of the formation of several H bonds.

Energy barriers for the addition of the $\cdot\text{OH}$ on the caffeic acid are found all below the asymptote, suggesting that this pathway occurs very easily. Exceptions to it are represented by the $\cdot\text{OH}$ addition on the C_1 and C_3OOH positions for which TSs are computed to lie at 1.64 and 9.33 kcal/mol, at the M05-2X level, and 0.18 and 8.04 kcal/mol, at the B3LYP level, over the asymptote. This is quite expected since $\cdot\text{OH}$ addition to the C_1 site entails the rupture of the π delocalization from the aromatic ring to the $-\text{CH}=\text{CH}-\text{COOH}$ chain and the rearrangement of the latter, while RAF at the C_3OOH destroys the delocalization involving the carboxylic group locating the unpaired electron to the oxygen in the carbonyl.

The final adducts are found to lie below the asymptote for all examined cases, indicating that the RAF process is quite

thermodynamically favored. The P's for the RAF $\text{C}_{1'}$ and RAF $\text{C}_{2'}$ are found at -35.41 and -36.48 kcal/mol, at the M05-2X level, and -28.38 and -30.27 kcal/mol, at the B3LYP level. In these adducts, the unpaired electron may be delocalized on the oxygen in the carbonyl in the case of RAF $\text{C}_{2'}$ and over the aromatic ring in the case of the RAF $\text{C}_{1'}$, as suggested by the atomic spin densities reported in Table S2 in the Supporting Information for the final adducts. P for the RAF C_4 is found at -34.49 (M05-2X) and -29.06 (B3LYP) kcal/mol. The odd electron may be delocalized over the aromatic ring and up to the $\text{C}_{1'}$ and $\text{C}_{2'}$ carbon atoms, even if no involvement of the carboxylic group is detected in this case (see atomic spin densities in Table S2). This significant possibility of delocalization of the unpaired electron in the final adduct, which does not occur for the RAF at the remaining positions, is the main feature of stabilization.

Considering both the thermodynamics and the kinetics of the process, the most reactive site within the RAF process is represented by the C_4 position. In fact, the lowest barrier corresponds to the radical addition to this site, as well as a good thermodynamic stabilization. This is mainly due to the broad electronic delocalization that occurs when $\cdot\text{OH}$ adds to the C_4 carbon atom. However, the radical addition to all carbon atoms, with the only exception being the carbon in the $-\text{COOH}$ group, seems to be quite viable since all stationary points are predicted to lie below the entrance channel.

A single electron transfer (SET) process occurs without passing through a saddle point, so for this channel, information about the structure of the radical cation of caffeic acid upon single electron removal and reaction energy (ΔE) for the process are provided.

The caffeic acid radical cation is a planar species as the parent molecule, at both the M05-2X and B3LYP levels. H bonds are retained upon electron removal (Figure 5). Analysis of the atomic spin densities (Table S3 in the Supporting Information) reveals that the unpaired electron is delocalized over the aromatic ring, with the exception of the C_5 carbon atom, and also on the $-\text{CH}=\text{CH}-\text{COOH}$ chain, especially involving $\text{C}_{1'}$ and $\text{C}_{2'}$.

The ΔE of reaction for the SET process is computed to be 149.38 and 140.22 kcal/mol, at the M05-2X and B3LYP levels, so it is found to be endergonic by both functionals. So, this mechanism may be ruled out for the possible radical scavenging activity of caffeic acid toward the hydroxyl radical with respect to HAT and RAF pathways.

Concerning the performance of the two exchange-correlation functionals for the prediction of the three mechanisms, we may point out that both give shapes for the classical reaction profiles that are qualitatively similar. But, they differ considerably in the values for the relative energies. In particular, B3LYP underestimates the barrier heights with respect to M05-2X. In some cases, B3LYP fails in locating RC and TS on the potential energy profile (C_3OH and C_4OH HAT, and C_3 and C_4 RAF). This behavior is mainly ascribed to its development since it has only been parametrized against a data set of thermochemistry, so its tendency toward predicting lower barrier heights is shown.¹⁰¹ Indeed, M05-2X has been parametrized and tested also against 38 barrier heights for hydrogen transfer (HT) processes, 18 of which involve radicals as reactants and products, and it is recommended by the authors to be used for thermochemical kinetics, noncovalent interactions (especially weak interaction, hydrogen bonding, $\pi-\pi$ stacking, and interactions energies of nucleobases), giving the best composite results for energetics.⁶⁶

Table 3. M05-2X/6-31+G(d) Activation (ΔE^\ddagger) and Reaction (ΔE) Energies (in kcal/mol) in Water and Benzene Media

	ΔE^\ddagger		ΔE	
	water	benzene	water	benzene
	SET			
	/	/	25.94	79.90
	HAT			
C ₂	7.57	6.78	-0.65	0.57
C ₅	6.47	4.75	-0.84	-0.39
C ₆	8.90	8.65	-1.08	-0.03
C _{1'}	4.47	3.35	-7.83	-6.43
C _{2'}	7.29	6.80	-2.18	-1.08
C ₃ OH	4.37	1.82	-36.20	-35.96
C ₄ OH	-0.85	0.30	-38.44	-38.30
C _{3'} OOH	0.95	0.24	-4.13	-3.28
	RAF			
C ₁	0.13	1.08	-16.40	-16.36
C ₂	-3.15	-2.63	-27.09	-23.99
C ₃	-2.00	-1.79	-25.44	-25.30
C ₄	-3.89	-4.74	-32.84	-34.13
C ₅	-1.55	-3.39	-22.03	-23.75
C ₆	-2.52	-2.39	-26.13	-25.38
C _{1'}	-2.99	-3.29	-34.31	-35.25
C _{2'}	-4.41	-4.58	-36.65	-36.92
C _{3'} OOH	10.56	9.54	-3.59	-4.56

Furthermore, it has been shown elsewhere¹⁰² that functionals that perform well for hydrogen-transfer barrier heights also perform well for barrier heights of more general classes of reaction.

Considering these arguments and also that M05-2X functionals have been successfully used by independent authors,^{71,74,77} we believe that the energetics predicted by M05-2X may be more reliable.

In Table 3, the activation (ΔE^\ddagger) and reaction (ΔE) energies are listed for all HAT, RAF, and SET mechanisms, in both polar (water) and nonpolar (benzene) media, to observe the effects of the environment on the energetics and thus on the viability of the three mechanisms investigated. The results are reported only with the M052X functional because of its better reliability, as discussed previously.

The channel requiring the lower activation energy within the HAT mechanism is the C₄OH one in a water medium (-0.85 kcal/mol), while in a benzene medium the C_{3'}OOH HAT presents the lowest barrier height (0.24 kcal/mol), even if C₄OH HAT follows with a barrier of 0.30 kcal/mol.

As far as the RAF process is concerned, the reactivity order in terms of activation energy in a water medium is C_{2'} (-4.41 kcal/mol) > C₄ (-3.89 kcal/mol) > C₂ (-3.15 kcal/mol) > C_{1'} (-2.99 kcal/mol) > C₆ (-2.52 kcal/mol) > C₃ (-2.00 kcal/mol) > C₅ (-1.55 kcal/mol) > C₁ (0.13 kcal/mol) > C_{3'} (10.56 kcal/mol). In a benzene medium, the order found is C₄ (-4.74 kcal/mol) > C_{2'} (-4.58 kcal/mol) > C₅ (-3.39 kcal/mol) > C_{1'} (-3.29 kcal/mol) > C₂ (-2.63 kcal/mol) > C₆ (2.39 kcal/mol) > C₁ (1.08 kcal/mol) > C_{3'} (9.54 kcal/mol). So, the medium does not seem to influence the barrier height or the reactivity order.

Also, the thermodynamics seem to be insensitive to both polar and nonpolar media since reaction energies are more or less the same computed in the gas phase. Again, HAT at the C₄OH and RAF at the C₄ positions are the more reliable pathways.

The SET pathway is likely to be favored by polar environments that may solvate the ionic species formed during the reaction, so the ΔE of the reaction in solution could provide a more complete investigation. In water and benzene, the ΔE values of reaction reported in Table 3 are quite affected by the solvent, and in particular, a water medium lowers the value by around 125 kcal/mol. The effect of a nonpolar environment is also quite important since the ΔE of reaction decreases by ~70 kcal/mol. However, the SET process remains to be thermodynamically unfavored with respect to HAT and RAF, also taking into account the solvent, so that it may be ruled out for the possible radical scavenging activity of caffeic acid toward the hydroxyl radical.

On the basis of these results, the HAT and RAF mechanisms are found to be very feasible as possible channels for the scavenging activity of the caffeic acid toward the hydroxyl radical. In the case of the former, the channel exhibiting the more reliable energetics is the hydrogen abstraction from the hydroxyl attached at the C₄. The RAF process indeed provides very low activation energies and thermodynamical feasibility regardless of the particular site of addition, with a slight preference for the insertion at the C₄ position. The presence of the environment and its polarity do not seem to particularly influence the energetics involved.

3.2. OH Scavenging by Caffeic Acid in the Deprotonated form. The pK_a of caffeic acid is 4.62,⁵⁸ indicating that this compound will primarily exist in anion form in the carboxylic functionality in the environment at physiological pH. So, to provide a detailed exploration about the radical scavenging activity of caffeic acid toward \bullet OH, also the anionic form of caffeic acid has been taken into account, considering HAT, RAF, and SET mechanisms.

The anion of the caffeic acid is also a planar species, at both the M05-2X and B3LYP levels.

Concerning the HAT mechanism, the RC stationary points are characterized by the formation of strong H bonds involving the \bullet OH radical and the carboxylate group of the caffeic acid. The only exception is the HAT at the C₂H position, in which the hydrogen-like interaction is established with the 3OH group at both the M05-2X and B3LYP levels (see Figures S1 and S2 in the Supporting Information). HAT at the C₂H proceeds toward the transition state, in which the H is shared between the OH radical (C₂H...OH distance is 1.24 and 1.20 Å) and the carbon atom (C₂...HOH distance is 1.25 and 1.35 Å), leading to the product complex (see Figures S1 and S2).

Both functionals fail in locating, on the potential energy profile, the reactant complex coming from the interaction between the \bullet OH and the hydrogen attached to the C₅, C₃OH, and C₄OH positions. Concerning the two latter cases, the optimizations yield as final results the product complex for the HAT C₃OH and HAT C₄OH, respectively. In the former, indeed, M05-2X and B3LYP computations yield the product complex PC referable to the HAT C₄OH; i.e., no stationary points (RC, TS, PC, and thus P) for the HAT C₅ are found, so this channel is ruled out.

HAT at the C₆H site starts with the formation of a reactant complex characterized by a strong H bond (1.63 Å at both M05-2X and B3LYP) between the hydroxyl radical and one oxygen atom of the carboxylate. Through a transition state

(1511i (M05-2X) and 1630i (B3LYP) cm^{-1}) in which the hydrogen is contemporaneously bonded to both the C_6 atom (1.23 (M05-2X) and 1.33 (B3LYP) Å) and the $\bullet\text{OH}$ group (1.27 (M05-2X) and 1.20 (B3LYP) Å), it is transferred to the hydroxyl radical forming the product complex between the H_2O and the C_6 radicalized caffeic acid.

The H-atom transfer reaction occurring between the $\bullet\text{OH}$ radical and the C_1H sees the formation of a reactant complex obtained for a $\text{C}_1\text{H}\cdots\text{OH}$ distance of 1.63 Å at both levels of theory. The transition state, necessary to proceed toward the product complex, is characterized by a $\text{C}_1\cdots\text{HOH}$ and a $\text{C}_1\text{H}\cdots\text{OH}$ distance of 1.24 and 1.25 Å at the M05-2X (imaginary frequency is 1498 cm^{-1}) level and 1.29 and 1.21 Å at the B3LYP level (imaginary frequency is 1534 cm^{-1}). Both M05-2X and B3LYP product complexes predict the delocalization of the unpaired electron on the whole molecule with the exception of the carboxylate, which is quite perpendicular to the rest of the molecule (see Figures S1 and S2, Supporting Information).

For the HAT process at the C_2H position, the reactant complex shows a H bond established between the hydroxyl radical and the carboxylate group (1.60 and 1.62 Å, at M05-2X and B3LYP, respectively). Both M05-2X and B3LYP saddle points (imaginary frequencies are 1624 and 1159 cm^{-1}) for the H-transfer show the incoming bond of the hydrogen to the $\bullet\text{OH}$ ($\text{C}_2\cdots\text{H}$ and $\text{C}_2\text{H}\cdots\text{OH}$ distances are 1.21 and 1.31 Å and 1.25 and 1.29 Å, at the M05-2X and B3LYP levels, respectively). In the product complex, the conjugation between the $-\text{CH}=\text{CH}-\text{COO}^-$ and the aromatic ring is lost, especially as far as the B3LYP computations are concerned ($\text{O}-\text{C}_3'-\text{C}_2'-\text{C}_1'$ dihedral is 38.81°).

Product complexes for the C_3OH and C_4OH HAT show the water molecule hydrogen bonded to the radicalized oxygen atom in the caffeic acid. B3LYP indicates that the carboxylate group is not coplanar with the rest of the molecule, as can be noted from Figures S1 and S2 (Supporting Information).

With regard to the energetics of the HAT process for the anion of the caffeic acid (values in parentheses in Table 1), we may conclude again that the hydrogen attached to the C_4 carbon atom is the most susceptible to be transferred to the hydroxyl radical, in such a way that the computations do not find a complex between reactants, and the thermodynamics are particularly favored in this case.

As expected, B3LYP underestimates barrier heights and overestimates thermodynamics of all channels of the HAT process with respect to the M05-2X.

RAF at the C_1 starts with the formation of the reactant complex that again is characterized by the establishment of a H bond between the hydroxyl radical and the carboxylate group, as occurred for the HAT at the C_6 , C_1' , and C_2' and as it will occur for the RAF at the C_2 , C_5 , C_6 , C_1' , and C_2' . In the saddle point for the $\bullet\text{OH}$ addition, whose nature of transition state is confirmed by the imaginary frequencies at 434 and 331 cm^{-1} , at the M05-2X and B3LYP levels, respectively, the distance $\text{HO}\cdots\text{C}_1$ assumes the values of 2.04 Å (M05-2X) and 2.00 Å (B3LYP). The final adduct shows a $\text{HO}-\text{C}$ σ bond and the disappearance of the conjugation between the $-\text{CH}=\text{CH}=\text{COO}-$ and the phenyl (see Figures S3 and S4 in the Supporting Information).

After the formation of the reactant complex, the RAF process at C_2 proceeds toward the final adduct after passing through the transition state (434i (M05-2X) and 236i (B3LYP) cm^{-1}) in which the $\bullet\text{OH}$ is approaching the C_2 carbon atom at 2.07 and 2.10 Å, at the M05-2X and B3LYP levels, respectively.

The reactant complex for the RAF C_3 presents a $\bullet\text{OH}$ radical pointing toward the carbon atom at a distance of 2.89 and 2.97 Å, which becomes 2.08 and 2.00 Å in the saddle point (imaginary frequencies are 279 and 334 cm^{-1}), and 1.43 and 1.45 Å in the final adduct, at the M05-2X and B3LYP levels, respectively.

Both M05-2X and B3LYP computations fail in locating the reactant complex for the RAF at C_4 , since optimizations yield the final adduct at C_4 for a $\text{HO}-\text{C}_4$ distance of 1.45 and 1.47 Å, respectively. The B3LYP final adduct presents a loss of conjugation between the side chain and the rest of the molecule with respect to the M05-2X one.

$\bullet\text{OH}$ addition at the C_5 site occurs through a saddle point in which the radical approaches the carbon at 2.11 Å with an imaginary frequency of 185 cm^{-1} , as indicated by the M05-2X computations, and 1.94 Å with an imaginary frequency of 342 cm^{-1} considering the B3LYP ones. The $\text{HO}-\text{C}_5$ bond is finally established in the final adduct at 1.46 and 1.48 Å, as far as M05-2X and B3LYP optimizations are concerned.

Upon the formation of the reactant complex, the RAF process at C_6 evolves toward the transition state in which the $\text{HO}\cdots\text{C}_6$ bond distances are 2.11 and 2.23 Å and the imaginary frequencies are 376 and 250 cm^{-1} , at the M05-2X and B3LYP levels, respectively. In the next product of the reaction, the $\text{HO}-\text{C}_6$ bond length is 1.45 (M05-2X) and 1.47 (B3LYP) Å.

RAF at C_1' and C_2' entails the formation of the reactant complex characterized by a H bond with the negative carboxylate group in both cases. The transition state for the hydroxyl radical addition occurs for a $\text{HO}\cdots\text{C}_1'$ and $\text{HO}\cdots\text{C}_2'$ distance of 2.11 and 2.18 Å, at the M05-2X level, and 2.16 and 2.36 Å, at the B3LYP one. The nature of the TS is confirmed by the imaginary frequencies at 410 (M05-2X) and 206 (B3LYP) cm^{-1} and 373 (M05-2X) and 140 (B3LYP) cm^{-1} , for the RAF C_1' and RAF C_2' , respectively. Geometries of the final adduct within RAF at C_1' and C_2' are predicted to be very similar by both M05-2X and B3LYP functionals (see Figures S3 and S4, Supporting Information).

Finally, the RAF process at the C_3' in the carboxylate entails the transition through the saddle point in which the hydroxyl radical is approaching the carbon atom at 1.77 and 1.71 Å, with an imaginary frequency of 436 and 509 cm^{-1} , at the M05-2X and B3LYP levels, respectively. The product shows the formation of a $\text{HO}-\text{C}_3'$ bond whose distance is 1.45 and 1.47 Å, as far as the M05-2X and B3LYP computations are concerned.

Formation of the reactant complexes is exothermic for all sites except for the C_3 one, because of the different arrangement of the adduct, as indicated before. Transition states are found at 12.30 (C_1), 5.09 (C_2), 5.58 (C_3), 0.54 (C_5), 4.04 (C_6), -2.87 (C_1'), -3.10 (C_2'), and 18.38 (C_3') kcal/mol with respect to the asymptote, and when compared to the values obtained for the neutral form of the caffeic acid they show a reduced reactivity toward the $\bullet\text{OH}$ scavenging. Also, the thermodynamics of the process are less favored with respect to the neutral form of the antioxidant. Again, the C_4 site may be recognized as the most reactive in the radical adduct formation mechanism (the results in parentheses in Table 2).

The ΔE values of 57.50 (M05-2X) and 35.77 (B3LYP) kcal/mol for the single electron transfer (SET) process indicate that this mechanism is quite unfavored also when the antioxidant is present in the ionic form, and thus it can be concluded that caffeic acid does not follow this reaction channel.

3.3. Rate Constants. The M05-2X/6-31+G(d) computed rate constants as a function of the temperature for the H atom abstraction from the C_4OH group and for the radical addition at

Table 4. Rate Constants ($\text{M}^{-1} \text{s}^{-1}$) Computed at M052X Level of Theory for the HAT Mechanism at the 4OH Site of Caffeic Acid

T (K)	k_{TST}	k_{CVT}	$k_{\text{CVT/SCT}}$
150	3.82×10^{11}	1.25×10^{10}	1.22×10^{10}
200	5.64×10^{10}	4.57×10^9	4.35×10^9
300	9.76×10^9	1.85×10^9	1.65×10^9
400	4.65×10^9	1.31×10^9	1.12×10^9
500	3.31×10^9	1.07×10^9	9.33×10^8
600	2.85×10^9	1.06×10^9	9.39×10^8

Table 5. Rate Constants ($\text{M}^{-1} \text{s}^{-1}$) Computed at M052X Level of Theory for the RAF Mechanism at the C_4 Site of Caffeic Acid

T (K)	k_{TST}	k_{CVT}	$k_{\text{CVT/SCT}}$
150	1.26×10^{14}	1.26×10^{14}	1.25×10^{14}
200	3.20×10^{12}	2.90×10^{12}	2.16×10^{12}
300	9.33×10^{10}	7.29×10^{10}	6.00×10^{10}
400	1.84×10^{10}	1.33×10^{10}	1.15×10^{10}
500	7.59×10^9	5.29×10^9	4.70×10^9
600	4.55×10^9	3.06×10^9	2.78×10^9

the C_4 position in the caffeic acid by the HO^\bullet radical are collected in Tables 4 and 5. Our choice to report only the MEP at the M05-2X level is due to the fact that this functional provides more reliable kinetics and thermodynamics than B3LYP, also according to previous works on this matter.^{95,96}

Concerning the H-atom transfer mechanism, all rate constants (k_{TST} , k_{CVT} , and $k_{\text{CVT/SCT}}$) are on the order of 10^9 at 300 K (Table 4), meaning that the reaction is very fast and diffusion-limited, and underlining the excellent antioxidant activity of caffeic acid toward the $\bullet\text{OH}$ radical.

It is worth noting that important variational effects (i.e., the ratio between variational CVT and conventional TST rate constants) are present in the hydrogen abstraction reaction under investigation. The differences between the conventional transition-state rate constants (k_{TST}) and the variational transition-state rate constants (k_{CVT}) are caused by these marked variational effects, which are more evident at low values of temperature. At all temperatures, the variational transition state location, specified by the value of $s^*\text{CVT}(T)$, moves toward the reactant side of the PES. For example, the value of $s^*\text{CVT}(T)$ is found at -0.1355 bohr at 300 K.

As can be noted by the $k_{\text{CVT/SCT}}$ values reported in Table 4, and due to the little significant barrier height for the H abstraction from the C_4OH group, tunneling plays a rather modest role in the dynamics of the HAT process.

The computed rate constants for the RAF process (Table 5) indicate that this channel is slightly favored with respect to the former. In fact, at 300 K, the rate constants are on the order of 10^{10} .

In this case, variational effects on computed rate constants are less significant. At 300 K, k_{TST} and k_{CVT} values are 9.33×10^{10} and $7.29 \times 10^{10} \text{ M}^{-1} \text{ s}^{-1}$, indicating that the classical potential energy contribution predominates over the entropy contribution. Again, tunneling contributions on thermal rate constants are quite negligible, as expected since the addition reaction is a barrierless reaction and involves the movement of heavy atoms.

The calculated rate constants for both abstraction and addition reactions show a negative temperature dependence and, consequently, a negative activation energy, which may be expected for a free radical reaction involving $\bullet\text{OH}$ as the most reactive among the ROS species.

Pulse radiolytic experiments about the $\bullet\text{OH}$ scavenging capability of caffeic acid, conducted in slightly alkaline solutions,¹⁰³ have yielded rate constants of $7.4 \times 10^8 \text{ M}^{-1} \text{ s}^{-1}$. Other literature values^{104,105} are $5.5 \pm 0.8 \times 10^9 \text{ M}^{-1} \text{ s}^{-1}$ and $3.24 \times 10^9 \text{ M}^{-1} \text{ s}^{-1}$. Experimentally, the HAT mechanism is suggested to be the essential channel, followed by caffeic acid, and thus is responsible for these rate constants.¹⁰³ At 300 K, the theoretical rate constant computed by us for the abstraction reaction is on the same order of magnitude as the experimental values.

4. CONCLUSIONS

In this study, we have performed theoretical calculations to investigate the efficiency of a naturally occurring antioxidant molecule, i.e., the caffeic acid, in scavenging the very damaging $\bullet\text{OH}$ free radical, which can arise in living organisms.

Three different pathways of reaction have been deeply analyzed, H-atom abstraction, radical addition, and single electron transfer. All possible sites of reaction have been investigated in order to obtain the more reliable pathways.

Both hydrogen abstraction and radical addition channels are very feasible, in terms of thermodynamics and kinetics, in the gas phase and in polar and nonpolar media. Within the former mechanism, the most reactive site is the phenolic OH attached to the C_4 carbon in the aromatic ring, since a strong exothermicity (≈ 40 kcal/mol) and a negative barrier height (-1 kcal/mol) are obtained within this channel. $\bullet\text{OH}$ addition on the C_4 carbon atom is the favorite within radical addition formation. The single electron transfer process seems to be very unlikely since unfavorable thermodynamics are found for this mechanism, regardless of the presence on the environment.

The rate constants for the two most probable mechanisms were independently calculated by using variational transition-state theory with multidimensional tunneling. Obtained values are 1.8548×10^9 and $7.2868 \times 10^{10} \text{ M}^{-1} \text{ s}^{-1}$, for the hydrogen abstraction and radical addition, respectively. Variational effects are significant in the case of the hydrogen transfer process, due to the flat potential energy profile, so that conventional transition state theory does not seem to be adequate for this reaction.

Tunneling seems to have a negligible effect in both cases.

We have found that the $\bullet\text{OH}$ addition mechanism on caffeic acid is slightly favored with respect to the hydrogen transfer.

The final rate constants present a negative temperature dependence, leading to a negative activation energy. This negative value may be due to the negative value of the enthalpy associated with the generalized transition state, where the tunneling effect is negligible.

The agreement with available experimental data lends confidence to the theoretical level and model used.

■ ASSOCIATED CONTENT

S Supporting Information. Figures showing the M05-2X and B3LYP optimized geometries of the stationary points coming from the reaction between the anion of caffeic acid and the hydroxyl radical, for HAT, RAF, and SET mechanisms. Tables reporting the atomic spin densities at the M05-2X level

in the products coming from HAT, RAF, and SET mechanisms. Figures showing the minimum energy path and the adiabatic potential-energy profiles computed at the M052X level for the HAT and RAF mechanisms. Complete refs 97 and 98. This material is available free of charge via the Internet at <http://pubs.acs.org>.

AUTHOR INFORMATION

Corresponding Author

*E-mail: nrusso@unical.it.

ACKNOWLEDGMENT

The University of Calabria and the Food Science & Engineering Interdepartmental Center of the University of Calabria, L.I.P.A.C., Calabrian Laboratory of Food Process Engineering (Regione Calabria APQ—Ricerca Scientifica e Innovazione Tecnologica I atto integrativo, Azione 2 laboratori pubblici di ricerca mission oriented interfiliere), and FONCICYT Mexico—EU "RMAYS" network, Project No. 94666, are gratefully acknowledged.

REFERENCES

- (1) Hertog, M. G. L.; Hollman, P. C. H.; Katan, M. B.; Kromhout, D. Intake of potentially anticarcinogenic flavonoids and their determinants in adults in the Netherlands. *Nutr. Cancer* **1993**, *20*, 21.
- (2) Robak, J.; Gryglewski, R. J. Bioactivity of flavonoids. *Pol. J. Pharmacol.* **1996**, *48*, 555.
- (3) Clifford, M. N. Chlorogenic acids and other cinnamates — nature, occurrence and dietary burden. *J. Sci. Food Agric.* **1999**, *79*, 362.
- (4) Rice-Evans, C.; Spencer, J. E.; Schroeter, H.; Rechner, A. R. Bioavailability of Flavonoids and Potential Bioactive Forms in Vivo. *Drug Metab. Drug Interact.* **2000**, *17*, 291.
- (5) Rencher, A. R.; Spencer, P. E.; Kuhnle, G.; Hahn, U.; Rice-Evans, C. A. Novel biomarkers of the metabolism of caffeic acid derivatives in vivo. *Free Radical Biol. Med.* **2001**, *30*, 1213.
- (6) Brit, D. F.; Hendrich, S.; Wang, W. Dietary agents in cancer prevention: flavonoids and isoflavonoids. *Pharmacol. Ther.* **2001**, *90*, 157.
- (7) Ross, J. A.; Kasum, C. M. DIETARY FLAVONOIDS: Bioavailability, Metabolic Effects, and Safety. *Ann. Rev. Nutr.* **2002**, *22*, 19.
- (8) Kris-Etherton, P. M.; Hecker, K. D.; Bonanome, A. B.; Coval, S. M.; Binkoski, A. E.; Hilpert, K.; Griel, A. E.; Etherton, T. D. Bioactive compounds in foods: their role in the prevention of cardiovascular disease and cancer. *Am. J. Med.* **2002**, *113*, 71.
- (9) Bravo, L. Polyphenols: Chemistry, Dietary Sources, Metabolism, and Nutritional Significance. *Nutr. Rev.* **1998**, *56*, 317.
- (10) Harborne, J. B. In *Plant Flavonoids in Biology and Medicine 11: Biochemical, Cellular and Medicinal Properties*; Alan R. Liss: New York, 1988; p 17.
- (11) Swain, T. In *Plant Flavonoids in Biology and Medicine: Biochemical, Pharmacological, and Structure-Activity Relationships*; Alan R. Liss: New York, 1986; p 1.
- (12) Harborne, J. B. In *Plant Flavonoids in Biology and Medicine: Biochemical Pharmacological, and Structure-Activity Relationships*; Alan R. Liss: New York, 1986; p 15.
- (13) Renaud, S.; de Lorgeril, M. Wine, alcohol, platelets, and the French paradox for coronary heart disease. *Lancet* **1992**, *339*, 1523.
- (14) Frankel, E. N.; Kanner, J.; German, J. B.; Parks, E.; Kinsella, J. E. Inhibition of oxidation of human low-density lipoprotein by phenolic substances in red wine. *Lancet* **1993**, *341*, 454.
- (15) Robbins, R. J. Phenolic acids in foods: an overview of analytical methodology. *J. Agric. Food Chem.* **2003**, *51*, 2866.
- (16) King, A.; Young, G. Characteristics and Occurrence of Phenolic Phytochemicals. *J. Am. Diet. Ass.* **1999**, *99*, 213.
- (17) Cody, V.; Middleton, E.; Harborne, J. B. In *Plant Flavonoids in Biology and Medicine-Biochemical, Pharmacological and structure-activity relationships*; Alan R. Liss: New York, 1986; p 429.
- (18) Cody, V.; Middleton, E.; Harborne, J. B.; Beretz, A. In *Plant Flavonoids in Biology and Medicine II-Biochemical, Cellular and Medicinal Properties*; Alan R. Liss: New York, 1988; p 187.
- (19) Das, N. P. In *Flavonoids in Biology and Medicine III-Current Issues in Flavonoids Research*; Singapore University Press: Singapore, 1990; p 213.
- (20) Kandaswami, C.; Middleton, E. Free radical scavenging and antioxidant activity of plant flavonoids. *Adv. Exp. Med. Biol.* **1994**, *366*, 351.
- (21) Duarte, J.; Perez-Vizcaino, F.; Utrilla, P.; Jimenez, J.; Tanargo, J.; Zarzuelo, A. Vasodilatory effects of flavonoids in rat aortic smooth muscle. Structure-activity relationships. *Gen. Pharmacol.* **1993**, *24*, 857.
- (22) Duarte, J.; Perez-Vizcaino, F.; Zarzuelo, A.; Jimenez, J.; Tanargo, J. Vasodilator effects of quercetin in isolated rat vascular smooth muscle. *Eur. J. Pharmacol.* **1993**, *239*, 1.
- (23) Brown, J. P. A review of the genetic effects of naturally occurring flavonoids, anthraquinones and related compounds. *Mut. Res.* **1980**, *75*, 243.
- (24) Mabry, T. J.; Markham, K. R.; Chari, V. M. In *The Flavonoids. Advances in Research*; Harborne, J. R., Mabry, T. J., Eds.; Chapman and Hall: London, 1982; p 52.
- (25) Robak, J.; Shridi, F.; Wolbis, M.; Krolikowska, M. Screening of the influence of flavonoids on lipoxygenase and cyclooxygenase activity. *Pol. J. Pharmacol. Pharm.* **1988**, *40*, 451.
- (26) Ho, C.-T.; Chen, Q.; Shi, H.; Zhang, K.-Q.; Rosen, R. T. Antioxidative effect of polyphenol extract prepared from various Chinese teas. *Prev. Med.* **1992**, *21*, 520.
- (27) Middleton, E.; Kandaswami, C. Effects of flavonoids on immune and inflammatory cell functions. *Biochem. Pharmacol.* **1992**, *43*, 1167.
- (28) Jovanovic, S. V.; Jankovic, I.; Josimovic, L. Electron-Transfer Reactions of Alkyl Peroxy Radicals. *J. Am. Chem. Soc.* **1992**, *114*, 9018.
- (29) Sogawa, S.; Nihro, Y.; Ueda, H.; Uzumi, A.; Miki, T.; Matsumoto, H.; Satoh, T. 3,4-Dihydroxychalcones as Potent 5-Lipoxygenase and Cyclooxygenase Inhibitors. *J. Med. Chem.* **1993**, *36*, 3904.
- (30) Lindahl, M.; Tagesson, C. Selective inhibition of group II phospholipase A₂ by quercetin. *Inflammation* **1993**, *17*, 573.
- (31) Elliot, A. J.; Scheiber, S. A.; Thomas, C.; Pardini, R. S. Inhibition of glutathione reductase by flavonoids: A structure-activity study. *Biochem. Pharmacol.* **1992**, *44*, 1603.
- (32) Chang, W. S.; Lee, Y. J.; Lu, F. J.; Chang, H. C. Inhibitory effects of flavonoids on xanthine oxidase. *Anticancer Res.* **1993**, *13*, 2165.
- (33) Report of the joint WHO/FAO expert consultation. Diet, nutrition and the prevention of chronic diseases. WHO Technical Report Series, No. 916 (TRS916), 2003.
- (34) de Groot, H. Reactive oxygen species in tissue injury. *Hepato-gastroenterol.* **1994**, *41*, 328.
- (35) Grace, P. A. Ischaemia-reperfusion injury. *Brit. J. Surg.* **1994**, *81*, 637.
- (36) Halliwell, B. How to characterize an antioxidant: an update. *Biochem. Soc. Symp.* **1995**, *61*, 73.
- (37) Darias-Martin, J.; Martin-Luis, B.; Carrillo-Lopez, M.; Lamuela-Raventos, R.; Diaz-Romero, C.; Boulton, R. Effect of Caffeic Acid on the Color of Red Wine. *J. Agric. Food Chem.* **2002**, *50*, 2062.
- (38) Gulcin, I. Antioxidant activity of caffeic acid (3,4-dihydroxycinnamic acid). *Toxicology* **2006**, *217*, 213.
- (39) Kroon, P. A.; Williamson, G. Hydroxycinnamates in plants and food: current and future perspectives. *J. Sci. Food Agric.* **1999**, *79*, 355.
- (40) Fukumoto, L. R.; Mazza, G. Assessing Antioxidant and Prooxidant Activities of Phenolic Compounds. *J. Agric. Food Chem.* **2000**, *48*, 3597.
- (41) Frank, H.; Thiel, D.; Macleod, J. Mass spectrometric detection of cross-linked fatty acids formed during radical-induced lesion of lipid membranes. *Biochem. J.* **1989**, *260*, 873.
- (42) Gebhardt, R.; Fausel, M. Antioxidant and hepatoprotective effects of artichoke extract and constituents in cultured rat hepatocytes. *Toxicology* **1997**, *11*, 669.
- (43) Joyeux, M.; Lobstein, A.; Anton, R.; Mortier, F. Comparative antilipoperoxidant, antinecrotic and scavenging properties of terpenes and biflavones from Ginkgo and some flavonoids. *Planta Med.* **1995**, *61*, 126.

- (44) Challis, B. C.; Bartlett, C. D. Possible cocarcinogenic effects of coffee constituents. *Nature* **1975**, *254*, 532.
- (45) Iwashii, H.; Ishii, T.; Sugata, R.; Kido, R. The effects of caffeic acid and its related catechols on hydroxyl radical formation by 3-hydroxyanthranilic acid, ferric chloride, and hydrogen peroxide. *Arch. Biochem. Biophys.* **1990**, *276*, 242.
- (46) Sudina, G. F.; Mirzoeva, O. K.; Pushkareva, M. A.; Korshunova, G. A.; Sumbatyan, N. V.; Varfolomeev, S. D. Caffeic acid phenethyl ester as a lipoxygenase inhibitor with antioxidant properties. *FEBS Lett.* **1993**, *329*, 21.
- (47) Rao, M. V.; Paliyath, G.; Ormrod, D. P. Ultraviolet-B- and Ozone-Induced Biochemical Changes in Antioxidant Enzymes of Arabidopsis thaliana. *Plant Physiol.* **1996**, *110*, 125.
- (48) Meyer, A. S.; Donovan, J. L.; Pearson, D. A.; Waterhouse, A. L.; Frankel, E. N. Fruit Hydroxycinnamic Acids Inhibit Human Low-Density Lipoprotein Oxidation in Vitro. *J. Agric. Food Chem.* **1998**, *46*, 1783.
- (49) Cartron, E.; Carbonneau, M. A.; Fouret, G.; Descomps, B.; Leger, C. L. Specific Antioxidant Activity of Caffeoyl Derivatives and Other Natural Phenolic Compounds: LDL Protection against Oxidation and Decrease in the Proinflammatory Lysophosphatidylcholine Production. *J. Nat. Prod.* **2001**, *64*, 480.
- (50) Kikuzaki, H.; Hisamoto, M.; Hirose, K.; Akiyama, K.; Taniguchi, H. Antioxidant Properties of Ferulic Acid and Its Related Compounds. *J. Agric. Food Chem.* **2002**, *50*, 2161.
- (51) Wright, J. S.; Johnson, E. R.; Di Labio, G. A. Predicting the Activity of Phenolic Antioxidants: Theoretical Method, Analysis of Substituent Effects, and Application to Major Families of Antioxidants. *J. Am. Chem. Soc.* **2001**, *123*, 1173.
- (52) Leopoldini, M.; Prieto Pitarch, I.; Russo, N.; Toscano, M. Structure, Conformation, and Electronic Properties of Apigenin, Luteolin, and Taxifolin Antioxidants. A First Principle Theoretical Study. *J. Phys. Chem. A* **2004**, *108*, 92.
- (53) Leopoldini, M.; Marino, T.; Russo, N.; Toscano, M. Density functional computations of the energetic and spectroscopic parameters of quercetin and its radicals in the gas phase and in solvent. *Theor. Chem. Acc.* **2004**, *111*, 210.
- (54) Leopoldini, M.; Marino, T.; Russo, N.; Toscano, M. Antioxidant Properties of Phenolic Compounds: H-Atom versus Electron Transfer Mechanism. *J. Phys. Chem. A* **2004**, *108*, 4916.
- (55) Jovanovic, S. V.; Steenken, S.; Simic, M. G.; Hara, Y. In *Flavonoids in Health and Disease*; Rice-Evans, C., Packer, L., Eds.; Marcel Dekker: New York, 1998; p 137.
- (56) Leopoldini, M.; Russo, N.; Chiodo, S.; Toscano, M. Iron Chelation by the Powerful Antioxidant Flavonoid Quercetin. *J. Agric. Food Chem.* **2006**, *54*, 6343.
- (57) Belyakov, V. A.; Roginsky, V. A.; Bors, W. Rate constants for the reaction of peroxy radical with flavonoids and related compounds as determined by the kinetic chemiluminescence method. *J. Chem. Soc., Perkin Trans. 2* **1995**, 2319.
- (58) Kono, Y.; Kobayashi, K.; Tagawa, S.; Adachi, K.; Ueda, A.; Sawa, Y.; Shibata, H. Antioxidant activity of polyphenolics in diets: Rate constants of reactions of chlorogenic acid and caffeic acid with reactive species of oxygen and nitrogen. *Biochim. Biophys. Acta* **1997**, *1335*, 335.
- (59) Foley, S.; Navaratnam, S.; McGarvey, D. J.; Land, E. J.; Truscott, T. G.; Rice-Evans, C. A. Singlet oxygen quenching and the redox properties of hydroxycinnamic acids. *Free Radical Biol. Med.* **1999**, *26*, 1202.
- (60) Lu, Y.; Foo, L. Y. Salvianolic acid L, a potent phenolic antioxidant from *Salvia officinalis*. *Tetrahedron Lett.* **2001**, *42*, 8223.
- (61) Lu, Y.; Foo, L. Y. Antioxidant activities of polyphenols from sage (*Salvia officinalis*). *Food Chem.* **2001**, *75*, 197.
- (62) Becke, A. D. Density-functional thermochemistry. III. The role of exact exchange. *J. Chem. Phys.* **1993**, *98*, 5648.
- (63) Lee, C.; Yang, W.; Parr, R. G. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B* **1988**, *37*, 785.
- (64) Becke, A. D. A new mixing of Hartree-Fock and local density-functional theories. *J. Chem. Phys.* **1993**, *98*, 1372.
- (65) Becke, A. D. Density-functional exchange-energy approximation with correct asymptotic behavior. *Phys. Rev. A* **1988**, *38*, 3098.
- (66) Zhao, Y.; Schultz, N. E.; Truhlar, D. G. Design of Density Functionals by Combining the Method of Constraint Satisfaction with Parametrization for Thermochemistry, Thermochemical Kinetics, and Noncovalent Interactions. *J. Chem. Theory Comput.* **2006**, *2*, 364.
- (67) Ditchfield, R.; Hehre, W. J.; Pople, J. A. Self-Consistent Molecular-Orbital Methods. IX. An Extended Gaussian-Type Basis for Molecular-Orbital Studies of Organic Molecules. *J. Chem. Phys.* **1971**, *54*, 724.
- (68) Hehre, W. J.; Ditchfield, R.; Pople, J. A. Self-Consistent Molecular Orbital Methods. XII. Further Extensions of Gaussian-Type Basis Sets for Use in Molecular Orbital Studies of Organic Molecules. *J. Chem. Phys.* **1972**, *56*, 2257.
- (69) Hariharan, P. C.; Pople, J. A. Accuracy of AH equilibrium geometries by single determinant molecular-orbital theory. *Mol. Phys.* **1974**, *27*, 209.
- (70) Gordon, M. S. The isomers of silacyclopropane. *Chem. Phys. Lett.* **1980**, *76*, 163.
- (71) Zavala-Oseguera, C.; Alvarez-Idaboy, J. R.; Merino, G.; Galano, A. OH Radical Gas Phase Reactions with Aliphatic Ethers: A Variational Transition State Theory Study. *J. Phys. Chem. A* **2009**, *113*, 13913.
- (72) Velez, E.; Quijano, J.; Notario, R.; Pabon, E.; Murillo, J.; Leal, J.; Zapata, E.; Alarcon, G. A computational study of stereospecificity in the thermal elimination reaction of menthyl benzoate in the gas phase. *J. Phys. Org. Chem.* **2009**, *22*, 971.
- (73) Vega-Rodriguez, A.; Alvarez-Idaboy, J. R. Quantum chemistry and TST study of the mechanisms and branching ratios for the reactions of OH with unsaturated aldehydes. *Phys. Chem. Chem. Phys.* **2009**, *11*, 7649.
- (74) Galano, A.; Alvarez-Idaboy, J. R. Guanosine + OH radical reaction in aqueous solution: A reinterpretation of the UV-Vis data based on thermodynamic and kinetic calculations. *Org. Lett.* **2009**, *11*, 5114.
- (75) Black, G.; Simmie, J. M. Barrier heights for H-atom abstraction by HO₂· from n-butanol—a simple yet exacting test for model chemistries? *J. Comput. Chem.* **2010**, *31*, 1236.
- (76) Furuncuoğlu, T.; Ugur, I.; Degirmenci, I.; Aviyyente, V. Role of Chain Transfer Agents in Free Radical Polymerization Kinetics. *Macromolecules* **2010**, *43*, 1823.
- (77) Galano, A.; Macías-Ruvalcaba, N. A.; Campos, O. N. M.; Pedraza-Chaverri, J. Mechanism of the OH Radical Scavenging Activity of Nordihydroguaiaretic Acid: A Combined Theoretical and Experimental Study. *J. Phys. Chem. B* **2010**, *114*, 6625.
- (78) Truhlar, D. G.; Kupperman, A. Exact tunneling calculations. *J. Am. Chem. Soc.* **1971**, *93*, 1840.
- (79) Fukui, K. Chemical reactivity theory - its pragmatism and beyond. *Pure Appl. Chem.* **1982**, *54*, 1825.
- (80) Miertus, S.; Scrocco, E.; Tomasi, J. Electrostatic interactions of a solute with a continuum. A direct utilization of ab initio molecular potentials for the prevision of solvent effects. *Chem. Phys.* **1981**, *55*, 117.
- (81) Miertus, S.; Tomasi, J. Approximate Evaluations of the Electrostatic Free Energy and Internal Energy Changes in Solution Processes. *Chem. Phys.* **1982**, *65*, 239.
- (82) Cossi, M.; Barone, V.; Cammi, R.; Tomasi, J. Ab initio study of solvated molecules: a new implementation of the polarizable continuum model. *Chem. Phys. Lett.* **1996**, *255*, 327.
- (83) Barone, V.; Cossi, M.; Menucci, B.; Tomasi, J. A new definition of cavities for the computation of solvation free energies by the polarizable continuum model. *J. Chem. Phys.* **1997**, *107*, 3210.
- (84) Garrett, B. C.; Truhlar, D. G. Criterion of minimum state density in the transition state theory of bimolecular reactions. *J. Chem. Phys.* **1979**, *70*, 1593.
- (85) Garrett, B. C.; Truhlar, D. G.; Grev, R. S.; Magnuson, A. W. Improved Treatment of Threshold Contributions in Variational Transition State Theory. *J. Phys. Chem.* **1980**, *84*, 1730.
- (86) Isaacson, A. D.; Truhlar, D. G. Polyatomic Canonical Variational Theory for Chemical Reaction Rates. Separable-Mode Formalism with Application to OH + H₂ → H₂O + H. *J. Chem. Phys.* **1982**, *76*, 1380.

(87) Truhlar, D. G.; Isaacson, A. D.; Garrett, B. C. In *Theory of Chemical Reaction Dynamics*; Baer, M., Ed.; CRC Press: Boca Raton, FL, 1985; Vol. 4, p 65.

(88) Fernandez-Ramos, A.; Miller, J. A.; Klippenstein, S. J.; Truhlar, D. G. Modeling of Bimolecular Reactions. *Chem. Rev.* **2006**, *106*, 4518.

(89) Liu, Y. P.; Lynch, G. C.; Truong, T. N.; Lu, D. H.; Truhlar, D. G.; Garrett, P. C. Molecular Modeling of the Kinetic Isotope Effect for the [1,5]-Sigmatropic Rearrangement of *cis*-1,3-Pentadiene. *J. Am. Chem. Soc.* **1993**, *115*, 2408.

(90) Lu, D. H.; Truong, T. N.; Melissas, V.; Lynch, G. C.; Liu, Y. P.; Garrett, B. C.; Steckler, R.; Isaacson, A. D.; Rai, S. N.; Hancock, G. C.; Laurderdale, J. G.; Joseph, T.; Truhlar, D. G. POLYRATE 4: A New Version of a Computer Program for the Calculation of Chemical Reaction Rates for Polyatomics. *Comput. Phys. Commun.* **1992**, *71*, 235.

(91) Truhlar, D. G.; Gordon, M. S. From Force Fields to Dynamics: Classical and Quantal Paths. *Science* **1990**, *249*, 491.

(92) Truong, T. N.; Lu, D. H.; Lynch, G. C.; Liu, Y. P.; Melissas, V.; Stewart, J. J. P.; Steckler, R.; Garrett, B. C.; Isaacson, A. D.; Gonzalez-Lafont, A.; Rai, S. N.; Hancock, G. C.; Joseph, T.; Truhlar, D. G. MORATE: A Program for Direct Dynamics Calculations of Chemical Reaction Rates by Semiempirical Molecular Orbital Theory. *Comput. Phys. Commun.* **1993**, *75*, 143.

(93) Espinosa-Garcia, J. Theoretical Study of the Trapping of the OOH Radical by Coenzyme Q. *J. Am. Chem. Soc.* **2004**, *126*, 920.

(94) Navarrete, M.; Rangel, C.; Espinosa-Garcia, J.; Corchado, J. C. Theoretical Study of the Antioxidant Activity of Vitamin E: Reactions of α -Tocopherol with the Hydroperoxy Radical. *J. Chem. Theory Comput.* **2005**, *1*, 337.

(95) Tejero, I.; Gonzalez-Garcia, N.; Gonzalez-Lafont, A.; Lluch, J. M. Tunneling in Green Tea: Understanding the Antioxidant Activity of Catechol-Containing Compounds. A Variational Transition-State Theory Study. *J. Am. Chem. Soc.* **2007**, *129*, 5846.

(96) Chiodo, S. G.; Leopoldini, M.; Russo, N.; Toscano, M. The inactivation of lipid peroxide radical by quercetin. A theoretical insight. *Phys. Chem. Chem. Phys.* **2010**, *12*, 7662.

(97) Frisch, M. J.; et al. *Gaussian 03*; Gaussian, Inc.: Pittsburgh PA, 2003.

(98) Corchado, J. C.; et al. *POLYRATE*, version 9.7; University of Minnesota, Minneapolis, MN, 2007.

(99) Zhao, Y.; Truhlar, D. G. Benchmark Databases for Nonbonded Interactions and Their Use To Test Density Functional Theory. *J. Chem. Theory Comput.* **2005**, *1*, 415.

(100) Leopoldini, M.; Russo, N.; Toscano, M. The molecular basis of working mechanism of natural polyphenolic antioxidants. *Food Chem.* **2011**, *125*, 288.

(101) Sousa, S. F.; Fernandes, P. A.; Ramos, M.-J. General Performance of Density Functionals. *J. Phys. Chem. A* **2007**, *111*, 10439 and references therein.

(102) Zhao, Y.; Gonzalez-García, N.; Truhlar, D. G. Benchmark Database of Barrier Heights for Heavy Atom Transfer, Nucleophilic Substitution, Association, and Unimolecular Reactions and Its Use to Test Theoretical Methods. *J. Phys. Chem. A* **2005**, *109*, 2012.

(103) Bors, W.; Michel, C.; Stettmaier, K.; Lu, Y.; Foo, L. Y. Pulse radiolysis, electron paramagnetic resonance spectroscopy and theoretical calculations of caffeic acid oligomer radicals. *Biochim. Biophys. Acta* **2003**, *1620*, 97.

(104) Li, X. F.; Cai, Z. L.; Katsumura, Y.; Wu, G. Z.; Muroya, Y. Reactions of reducing and oxidizing radicals with caffeic acid: a pulse radiolysis and theoretical study. *Radiat. Phys. Chem.* **2001**, *60*, 345.

(105) Wang, W. F.; Luo, J.; Yao, S. D.; Lian, Z. R.; Zhang, J. S.; Lin, N. Y.; Fang, R. Y.; Hu, T. X. Interaction of phenolic antioxidants and hydroxyl radicals. *Radiat. Phys. Chem.* **1993**, *42*, 985.