JCTC Journal of Chemical Theory and Computation

A Novel Hamiltonian Replica Exchange MD Protocol to Enhance Protein Conformational Space Sampling

Roman Affentranger,[†] Ivano Tavernelli,[‡] and Ernesto E. Di Iorio*[,†]

Institut für Biochemie, Eidgenössische Technische Hochschule ETH-Zurich, Schafmattstrasse 18, 8093 Zurich, Switzerland, and Institut de Chimie Moléculaire et Biologique, BCH-LCBC, Ecole Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland

Received October 13, 2005

Abstract: Limited searching in the conformational space is one of the major obstacles for investigating protein dynamics by numerical approaches. For this reason, classical all-atom molecular dynamics (MD) simulations of proteins tend to be confined to local energy minima, particularly when the bulk solvent is treated explicitly. To overcome this problem, we have developed a novel replica exchange protocol that uses modified force-field parameters to treat interparticle nonbonded potentials within the protein and between protein and solvent atoms, leaving unperturbed those relative to solvent−solvent interactions. We have tested the new protocol on the 18-residue-long tip of the P domain of calreticulin in an explicit solvent. With only eight replicas, we have been able to considerably enhance the conformational space sampled during a 100 ns simulation, compared to as many parallel classical molecular dynamics simulations of the same length or to a single one lasting 450 ns. A direct comparison between the various simulations has been possible thanks to the implementation of the weighted histogram analysis method, by which conformations simulated with modified force-field parameters can be assigned different weights. Interatom, inter-residue distances in the structural ensembles obtained with our novel replica exchange approach and by classical MD simulations compare equally well with those derived from NMR data. Rare events, such as unfolding and refolding, occur with reasonable statistical frequency. Visiting of conformations characterized by very small Boltzmann weights is also possible. Despite their low probability, such regions of the conformational space may play an important role in the search for local potential-energy minima and in dynamically controlled functions.

1. Introduction

Proteins are complex systems characterized by very rough free-energy landscapes (FEL). A feature that certainly contributes to complexity is the presence of anisotropic interactions—both within the protein and between the macromolecule and the surrounding solvent—where the coexistence of repulsive and attractive terms leads to many degenerate local energy minima. Such minima are separated by free-energy barriers, whose heights are often much larger than the thermal energy available to the system. For this reason, conventional all-atom molecular dynamics (MD) simulations of proteins in explicit solvent at room temperature suffer of a problem known as kinetic trapping; namely, the system tends to remain confined within one of the many local energy minima. Therefore, physical quantities that depend on an extensive sampling of the conformational space cannot be adequately calculated. Furthermore, conformations with very small Boltzmann weights, which are likely to be

* Corresponding author tel.: +41-44-6323137; fax: +41-44-6321298; e-mail: diiorio@bc.biol.ethz.ch.

[†] Eidgenössische Technische Hochschule ETH−Hönggerberg.

[‡] Ecole Polytechnique Fédérale de Lausanne.

10.1021/ct050250b CCC: $33.50
Published on Web 01/13/2006

involved in processes of high biological relevance, such as conformational transitions and dynamically controlled functional events,[1] are not visited.

A possible remedy to this problem is to perform MD simulations in a generalized ensemble (for a review, see ref 2). The idea is to achieve a random walk in potential-energy space, which allows the system to easily overcome the energy barriers that separate local minima and, therefore, to sample a much wider phase space compared to conventional simulations. Three well-known approaches for carrying out generalized ensemble MD simulations are the multicanonical algorithm,[3,4] simulated tempering,[5,6] and the replica exchange method (REM).[2,7−15] The former two algorithms make use of non-Boltzmann probability weight factors, which are not known a priori and need to be determined by trial simulations. This process is highly nontrivial and can be very tedious for complex systems such as proteins. In contrast, REM uses standard Boltzmann weight factors that are known a priori. A number of noninteracting simulations of the same system are performed in parallel, but under different conditions; at given time intervals, the simulation conditions are exchanged with a specific transition probability between replica pairs. Therefore, REM is particularly well-suited for parallel computing on simple PC clusters because it requires very little communication between the individual processors. The algorithm was originally developed for Monte Carlo simulations[15] and has been adapted to MD simulations by Sugita and Okamoto.[7] In its original implementation, the condition to be varied and exchanged among the replicas is the temperature. This results in a random walk in temperature space, which in turn induces a random walk in potential-energy space. Thus, systems that—when simulated by conventional methods at room temperature—would remain trapped within a limited region of the conformational space are allowed to escape more easily from local minima by jumping back and forth between high and low temperatures.

For large systems, such as proteins in an explicit solvent, temperature replica exchange MD (T-REMD) simulations have one major drawback: since the number of replicas needed to cover a given temperature range is roughly proportional to the square root of the number of degrees of freedom of the system,[16] many replicas need to be simulated, thus, rendering T-REMD simulations of proteins in an explicit solvent very demanding in computational terms. Because efficient sampling requires diffusion in temperature space, the higher the number of replicas that are used, the longer the simulation has to be performed, or the more frequently exchanges have to be attempted. This limits the capability of the method to obtain—with equivalent computational effort—better thermodynamic sampling, compared to classical MD (CMD) simulations. A remedy to this shortcoming of T-REMD is given by the Hamiltonian REM,[9,16,17] where the various replicas are simulated at constant or variable temperatures, but with different parameter sets for the equations of motion. This approach rests on the consideration that, since the individual simulations are independent and noninteracting, they need not necessarily be simulated using the same Hamiltonian. By restricting the changes introduced in the different Hamiltonians to only a subset of the degrees of freedom of the system, the number of replicas needed to cover a given range in "effective temperature" can be greatly reduced compared to T-REMD simulations. Both standard T-REMD and Hamiltonian REMD (H-REMD) at constant temperature are, thus, one-dimensional formulations of the general REMD methodology.[9] An obvious advantage of H-REMD at constant temperature is that no velocity rescaling[7] is needed when exchanges between replicas take place. Additional details on replica exchange approaches are summarized in a recent review by Snow et al.[18]

There is evidence that the dynamic properties of a protein are influenced by the frustration of its interparticle nonbonded interactions.[19−21] Therefore, H-REMD simulations, using modified force-field parameters for such interactions, are expected to enhance the sampling of the conformational space by directly influencing the frustration of the system and, therefore, its dynamic properties.

We report here the implementation of a new H-REMD protocol based on the simultaneous modification of electrostatic and Lennard-Jones (LJ) parameters, aiming also at testing this working hypothesis. Although not widespread in combination with H-REMD, the idea of modifying the force-field parameters used for classical MD simulations has already been exploited for locating the global minimum of the complex potential-energy hypersurface of oligopeptides.[22,23] Subsequently, this approach has been further developed to simulate, for instance, protein folding[24] or to predict transmembrane helix packing.[25] We have used our H-REMD to simulate—in an explicit solvent—the 18-residue-long tip of the P domain of calreticulin, a chaperon involved in protein quality control in the endoplasmic reticulum.[26] Recently, the structure of the long, flexible, hairpin-like P domain of calreticulin was solved by NMR, along with those of increasingly smaller fragments of its tip. All the fragments are shown to adopt the same structure they do in the full-length hairpin.[27−29] Now, also, the 18-residue-long polypeptide corresponding to the very tip of the calreticulin P domain has been investigated, but the data leave room for interpretation concerning its proteinlike folding behavior (L. Ellgaard, Institute of Biochemistry, ETH-Zurich, personal communication). Therefore, we used CRT18 as a test molecule for our new H-REMD protocol with the hope of gaining new insights on its folding properties.

2. Methods and Analysis

2.1. Simulation Details. All simulations were performed with the software package GROMACS 3.1.4,[30,31] using the GROMOS 43a1 force field[32] and periodic boundary conditions. Temperature (T) and pressure (P) were held constant using the weak coupling method,[33] with relaxation times of 0.1 ps for T and 0.5 ps for P. The protein and solvent, including ions, were each coupled separately to a temperature bath. All simulations were performed at 278 K under a pressure of 1 bar. An integration step of 2 fs was used, keeping bond lengths constant using the LINCS[34] algorithm for the peptide and SETTLE[35] for the water molecules. Nonbonded interactions were treated with the twin-range method, with cutoff radii of 0.8 and 1.4 nm, updating the

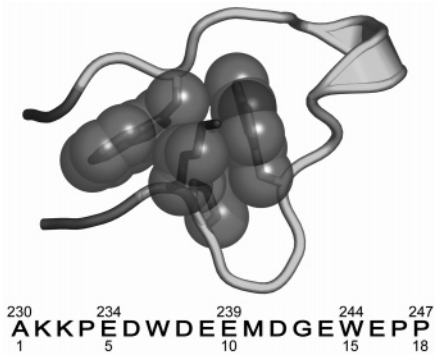


Figure 1. Structure of CRT18, as obtained from the coordinates of the larger 36 amino acid fragment (PDB entry 1K91), along with its amino acid sequence with the numbering pertaining to the full P domain of calreticulin (above the sequence), and that used in this work. This structure has been used as the starting configuration for all of our simulations and as reference for the analysis of the trajectories. The model highlights various regions of the molecule. Thus, the backbone is shown as a tube, with the first and last two residues in dark gray; the side chains of the amino acids involved in the formation of the hydrophobic core as sticks, with the individual atoms that make up the core represented as transparent spheres; and the single α-helical turn, formed by residues 9−12, as a ribbon. This region is recognized by DSSP[46] as being α-helical in only 2 of the 20 structures deposited in the PDB file 1K91. The model was drawn using PYMOL (http://www.pymol.org).

pair lists every step or every five steps respectively for the short and long cutoff. Long-range electrostatic interactions were treated with the reaction-field approach,[36] using a cutoff radius of 1.4 nm and a dielectric constant for the reaction field of 68.[37]

Initial coordinates for the 18-residue tip of the calreticulin P domain (see Figure 1) were obtained from the larger, 36 amino acid fragment (PDB entry 1K91[27]). The system was solvated in an octahedral box with 2808 SPC/E[38] water molecules, leaving an initial minimum distance between the peptide and the box walls of 1.1 nm. Aliphatic hydrogen atoms were treated by the united-atoms approach. Acidic residues were assumed to be in the charged state corresponding to neutral pH, leading to a net charge of −6. After energy minimization, the system was neutralized by adding eight sodium and two chlorine ions using the GROMACS program GENION. Following energy minimization of the neutralized system, the atoms were assigned random velocities drawn from a Maxwell distribution corresponding to 213 K. The system was then gradually heated to 278 K within 100 ps, applying a decreasing positional restraint to the protein atoms with force constants ranging between 25 000 and 0 kJ mol^{-1} nm^{-2}. From this state, a 450-ns CMD simulation was performed (long CMD, LCMD) on a dual processor Pentium III (1266 MHz) computer. In addition, eight independent, 100-ns-long, CMD simulations (8CMD) were carried out under the same conditions, assigning random initial velocities to each of them, followed by warming to 278 K using a decreasing positional restraint. These simulations were performed on a small cluster of single-processor Pentium IV (3.0 GHz) computers. For both the LCMD and each of the 8CMD simulations, protein coordinates were saved every 0.2 ps and the initial 25 ns were omitted for analysis, unless otherwise stated.

2.2. Hamiltonian Replica Exchange. For a system composed of N atoms with coordinate vectors and momentum vectors denoted respectively by $\mathbf{q} \equiv \{q_1, ..., q_N\}$ and $\mathbf{p} \equiv \{p_1, ..., p_N\}$, the Hamiltonian is the sum of the kinetic energy $K(\mathbf{p})$ and the potential energy $E(\mathbf{q})$:

$$H(\mathbf{q},\mathbf{p}) = K(\mathbf{p}) + E(\mathbf{q}) \tag{1}$$

Let us now consider one step of a simple H-REMD simulation at constant inverse temperature $\beta = 1/k_B T$ on two replicas i and j, with coordinate vectors $\mathbf{q}_i$ and $\mathbf{q}_j$ and momentum vectors $\mathbf{p}_i$ and $\mathbf{p}_j$, simulated respectively with the two Hamiltonians H_m and H_n, which differ only in their form of the potential energy:

$$H_k(\mathbf{q},\mathbf{p}) = K(\mathbf{p}) + E_k(\mathbf{q}) \qquad (k = m \text{ or } n) \tag{2}$$

This corresponds to a state Ω in the generalized ensemble

$$\Omega = \{H_m(\mathbf{q}_i,\mathbf{p}_i), H_n(\mathbf{q}_j,\mathbf{p}_j)\} \tag{3}$$

An exchange of Hamiltonians between the two replicas can be described as

$$\Omega = \{H_m(\mathbf{q}_i,\mathbf{p}_i), H_n(\mathbf{q}_j,\mathbf{p}_j)\} \rightarrow \Omega' = \{H_n(\mathbf{q}_i,\mathbf{p}_i), H_m(\mathbf{q}_j,\mathbf{p}_j)\} \tag{4}$$

In order for this exchange process to converge toward an equilibrium distribution, the condition of detailed balance must be imposed on the exchange probability $w(\Omega \rightarrow \Omega')$[15], leading to

$$w(\Omega \rightarrow \Omega') \equiv \min[1, \exp(-\Delta)] \tag{5}$$

with

$$\Delta \equiv \beta[E_n(\mathbf{q}_i) - E_m(\mathbf{q}_i) + E_m(\mathbf{q}_j) - E_n(\mathbf{q}_j)] \tag{6}$$

which is independent of the individual momenta of the two replicas.

Writing the potential-energy term of Hamiltonian k as

$$E_k(\mathbf{q}) = V_u(\mathbf{q}) + \sum_{l=1}^{L} \mu_{l,k} V_l(\mathbf{q}) \tag{7}$$

where $V_u(\mathbf{q})$ incorporates the unmodified part of the function and $\mu_{l,k}$ are the factors by which the $V_l(\mathbf{q})$ terms are scaled; Δ in eq 6 further reduces to

$$\Delta \equiv \beta \sum_{l=1}^{L} (\mu_{l,n} - \mu_{l,m})[V_l(\mathbf{q}_i) - V_l(\mathbf{q}_j)] \tag{8}$$

Thus, the exchange probability is only dependent on the modified part of the potential-energy function.

On the basis of the GROMOS 43a1 force-field and combination rules, we generated modified force fields multiplying the charges of side-chain protein atoms and the $C6^{1/2}$ and $C12^{1/2}$ LJ parameters of all protein atoms by a factor $f_k < 1$. The 1−4 LJ parameters, which are defined separately in the GROMOS force fields, were not scaled.

Hence, in our case, L is equal 2, and eq 7 becomes

$$E_k(\mathbf{q}) = V_u(\mathbf{q}) + f_k V_1(\mathbf{q}) + f_k^2 V_2(\mathbf{q}) \quad (9)$$

where V_1 represents the sum of all LJ protein−solvent interactions and electrostatic interactions between side-chain and main-chain atoms as well as between side-chain atoms and the solvent. V_2 accounts for the sum of LJ interactions between protein atoms and electrostatic interactions between side-chain atoms. Charges and LJ parameters of all solvent particles, including ions, were left unchanged. The spacing between factors f was chosen such that they would decrease roughly exponentially, and their exact values were tuned, by means of few short trial simulations, to yield exchange probabilities of roughly 20%. Seven modified force fields were thus generated, with f values of 0.965, 0.931, 0.898, 0.867, 0.837, 0.808, and 0.780. They were used, along with the unmodified GROMOS 43a1 force field, to run a 100 ns H-REMD simulation on eight replicas, starting from the same initial structure as that in the LCMD simulation, and attempting pairwise replica exchanges every 10 ps. Protein coordinates were saved every 50 fs, and the initial 25 ns of the simulation were omitted for analysis, unless otherwise stated. This simulation was run on a cluster consisting of eight dual-processor Pentium III (750 MHz) nodes.

2.3. Weighted Histogram Analysis Method. Data produced during REMD simulations can be combined using the weighted histogram analysis method (WHAM).[39,40] The algorithm was originally developed for umbrella-sampling simulations but can easily be adapted to our H-REMD approach by reformulating the potential-energy function corresponding to the ith force field in the following way:

$$E_i = E_0 + \lambda_{1,i} V_1 + \lambda_{2,i} V_2 \qquad \lambda_{1,i} = f_i - 1;\ \lambda_{2,i} = f_i^2 - 1 \quad (10)$$

where E_0 is the potential energy computed with the unmodified force field, while the scaling factor f_i and the potentials V_1 and V_2 are like those in eq 9. This form of the potential-energy function corresponds to that used in umbrella-sampling simulations.

Since we performed our H-REMD simulation at a constant temperature, the WHAM equations become independent of the value of E_0[41] and can be formulated in terms of the values of the biasing potentials V_1 and V_2 only:

$$\exp(-g_i) = \sum_{k=1}^{R}\sum_{t=1}^{n} \frac{\exp(-\beta\lambda_{1,i}V_{1,t}^{(k)} - \beta\lambda_{2,i}V_{2,t}^{(k)})}{\sum_{m=1}^{R} n \exp[g_m - \beta\lambda_{1,m}V_{1,t}^{(k)} - \beta\lambda_{2,m}V_{2,t}^{(k)}]} \quad (i = 1, ..., R) \quad (11)$$

where R is the number of replicas, n the number of snapshots used for the analysis, g_i a dimensionless free energy for the force field i, and $V_{(1,2),t}^{(k)}$ represents the value of the biasing potential for replica k at time t. After iterating the set of equations in eq 11 to self-consistency of the values of g_i, an un-normalized statistical weight P_0 is obtained for each time point t of each replica k, which has the form

$$P_0(k,t) = [\sum_{m=1}^{R} n \exp(g_m - \beta\lambda_{1,m}V_{1,t}^{(k)} - \beta\lambda_{2,m}V_{2,t}^{(k)})]^{-1} \quad (12)$$

and gives the probability of sampling point t of replica k with the unmodified force field.

2.4. Free-Energy Landscapes. As described in detail by Tavernelli et al.,[42] a two-dimensional representation of the FEL can be obtained from simulated atomic trajectories by plotting the negative logarithm of the joint probability distribution of two global parameters, ξ_1 and ξ_2. The resulting graph is a projection of the relative FEL, in units of k_BT, on a plane defined by the two global parameters. In the case of H-REMD simulations, the probability distribution must be calculated in a weighted manner; we did this using the WHAM-derived statistical weights.

2.5. Native Contacts. For the definition of a set of native contacts in CRT18, we have used the atomic trajectory between 250 and 750 ps of our LCMD simulation. Each residue was partitioned into main-chain and side-chain atoms, and the minimum interatomic distance between these groups was calculated for all snapshots in the trajectory, taking into account only amino acids at least three residues apart along the polypeptide chain. To be included in the list of native contacts, pairs had to display interatomic distances lower than 0.37 nm for at least 60% of the time. By this procedure, we have identified 25 native contacts, each of which was assigned a weight corresponding to its fractional presence during the 500 ps LCMD simulation segment. Summing up the weights of the resulting matrix, we have obtained a normalization factor F.

The existence of native contacts in all our simulations was monitored and scored by means of the matrix derived as just discussed. For each snapshot along a simulated trajectory, we defined the fraction of native contacts (FNC) as the sum of the weights-matrix elements corresponding to the native contacts present at that time, divided by the normalization factor F. Our analysis also included a monitoring of the total number of contacts as a function of time.

2.6. Clustering. We performed a structural clustering of the simulated ensembles using the algorithm described by Daura et al.[43] The procedure is based on the calculation of a matrix of the pairwise positional root-mean-square deviation (RMSD), after least-squares superposition, and the choice of a cutoff to define the neighborhood of each cluster. The structure with the largest number of neighbors is considered to be the center of the largest cluster, and it is removed, together with its neighbors, from the pool of structures before again searching for the next largest cluster. This procedure is repeated until all structures in the ensemble are clustered.

In our clustering analyses, we used a cutoff of 0.1 nm and considered only the backbone atoms of residues 3−16 (see Figure 1), both for the least-squares fitting and for the computation of the RMSD. To give different weights to the structures obtained with modified Hamiltonians during the H-REMD simulation, we did the clustering analysis using structures averaged over 5 ps intervals. For the classical MD

simulations, the weight of a cluster is equal to the number of its members divided by the total number of structures used for the clustering. In the case of the H-REMD simulation, the weight of a cluster equals the sum of the 5-ps-averaged WHAM weights of its members, normalized by the sum of the WHAM weights for the whole ensemble.

2.7. Correlation Coefficients. As a criterion to compare the convergence efficiency of the H-REMD to that of the 8CMD simulation, we have calculated the Pearson's correlation coefficients resting on 1D probability distributions of RMSD and FNC, WHAM-weighted in the case of H-REMD. This was done by computing the correlation coefficient $r(t)$, defined as

$$r(t) = \frac{[\sum_{i=1}^{N} P_t(i)\, P_0(i)] - N^{-1}\sum_{i=1}^{N} P_t(i) \sum_{i=1}^{N} P_0(i)}{\sqrt{\{[\sum_{i=1}^{N} P_t^{\,2}(i)] - N^{-1}[\sum_{i=1}^{N} P_t(i)]^2\}\{[\sum_{i=1}^{N} P_0^{\,2}(i)] - N^{-1}[\sum_{i=1}^{N} P_0(i)]^2\}}} \quad (13)$$

where $P_t(i)$ and $P_0(i)$ refer respectively to the probability distribution for bin i between 25 ns and time t and to that computed between 25 ns and 100 ns. Since the summations of $P_t(i)$ and $P_0(i)$ over all bins are equal to 1, eq 13 reduces to

$$r(t) = \frac{\sum_i P_t(i)\, P_0(i) - \frac{1}{N}}{\sqrt{\left\{[\sum_{i=1}^{N} P_t^{\,2}(i)] - \frac{1}{N}\right\}\left\{[\sum_{i=1}^{N} P_0^{\,2}(i)] - \frac{1}{N}\right\}}} \quad (14)$$

2.8. Comparison with NMR Data. To compare our simulations with experimental data, we used a set of upper bounds for the distances between pairs of hydrogen atoms belonging to different amino acids, as derived by nuclear Overhauser effect (NOE) measurements carried out by P. Bettendorff and L. Ellgaard (Institutes of Biochemistry and Molecular Biology and Biophysics, ETH-Zurich, unpublished results). The comparison was based on 126 unambiguous inter-residue interproton distances. These upper bounds were determined assuming the NOE intensity to be inversely proportional to the sixth power of the interproton distance. We therefore compared the simulated ensemble averages $\langle d^{-6}\rangle^{-1/6}$ (using WHAM weighting for the H-REMD simulation) with the upper bounds derived from NMR data.

3. Results and Discussion

The goal of REMD simulations is to enhance conformational space sampling compared to that of classical MD. The basic idea behind T-REMD simulations is to let the system experience elevated temperatures, thereby allowing it to more easily overcome high-energy barriers separating conformational states. Instead, in our H-REMD approach, we use different Hamiltonians with modified nonbonded interaction parameters, to enhance conformational space sampling not only by altering the height of energy barriers but also by affecting the frustration of the system and, therefore, its dynamic properties.[19] To our knowledge, this is the first report on a H-REMD simulation in an explicit solvent that also deals with a direct comparison to the results of the same number of classical MD simulations of identical length and those of a single, longer one. Previous H-REMD simulations have been carried out, for instance, by Fukunishi et al.[16] and Jang et al.[44] The former authors compared T-REMD to two variants of H-REMD, one using scaled hydrophobicity and the other phantom chains that allow various degrees of atomic overlaps and, therefore, the polypeptide chain to cross over itself. The simulations were carried out using a "coarse-grained" protein model in which (*i*) the solvent effect is implicitly accounted for via solvation free energy; (*ii*) the backbone includes three united atoms per amino acid, that is, NH, CH, and CO; and (*iii*) the side chains, except for glycine, are simplified as spheres placed at the center of mass of the residue. Using a 16-residue polyalanine and the albumin-binding domain of protein A, Fukunishi et al.[16] show that the scaled hydrophobicity method is most efficient. On the other hand, Jang et al.[44] used a generalized effective potential to achieve a change in the effective temperature of the system by modifying the torsional and nonbonded terms of the potential energy function. They carried out a 4.1 ns H-REMD simulation at 100 K (referred to as *q*-REM in their article), on an alanine dipeptide in vacuo, showing that two replicas, with q values of 1 and 1.002, are as good as at least five replicas in T-REMD of the same length, with temperatures of 100, 123, 148, 178, and 213 K.

3.1. H-REMD Protocol and WHAM. We tested several approaches to modify the individual terms describing the nonbonded interactions of a protein in water, namely, reducing only the LJ potentials between protein atoms, lowering exclusively the partial charges of side-chain atoms, or a combination of the two. The last approach (eq 9) gave the best results, both in terms of efficiency in sampling the conformational space and in terms of diffusion of the individual replicas in the space defined by the different Hamiltonians.

Crucial for the success of a replica exchange simulation is that (a) the exchanges occur frequently enough, such that each replica samples the whole range of the different conditions used for the simulation several times, and (b) the time gap between two exchange attempts is longer than the autocorrelation time of the potential energy. For our modified Hamiltonians, the autocorrelation time was determined to be in the sub-picosecond range. Therefore, one can assume a period of 10 ps between exchange attempts to be long enough to ensure quasi-independence of a replica-exchange step from the previous one.

As described in Section 2.2, we carried out short trial simulations to adjust the scaling factors f—used to prepare the modified force fields—to yield replica exchange probabilities of ~20%. This is neither a guarantee that the same probability levels are maintained for a longer simulation, where large conformational changes might influence them, nor does it ensure that each replica samples the whole range of simulation conditions. However, the data reported in Table 1 show that the average exchange probabilities between

Table 1. Percentages of Force Field Exchange during the H-REMD Simulation[a]

scaling factor	0.965	0.931	0.898	0.867	0.837	0.808	0.780
1.000	18.56	0.26	0.00	0.00	0.00	0.00	0.00
0.965		18.37	0.22	0.00	0.00	0.00	0.00
0.931			19.33	0.53	0.01	0.00	0.00
0.898				20.97	0.49	0.00	0.00
0.867					22.74	0.42	0.00
0.837						21.36	0.63
0.808							25.19

[a] Average exchange percentages between the individual force fields used in the H-REMD simulation computed from the entire trajectories. Force fields are represented by the scaling factors *f* used to modify the electrostatic and LJ parameters (see Section 2.2).

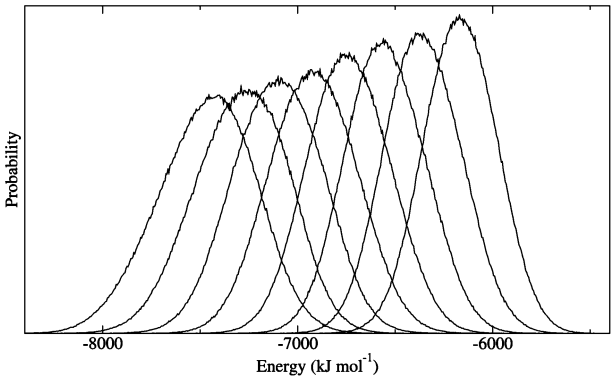


Figure 2. Histograms of the sum of the terms V_1 and V_2 in eq 9 for the individual force fields, as obtained from the whole 100 ns H-REMD simulation. The left-most and right-most curves correspond to the structural ensembles simulated respectively with the unmodified and most strongly modified force fields. The curves for neighboring force fields overlap considerably, ensuring sufficiently large replica-exchange probabilities.

neighboring force fields, when calculated for the whole simulation period of 100 ns, were sufficiently large. Furthermore, Figure 2 shows that neighboring histograms of the sums of the terms V_1 and V_2 in eq 9 display considerable overlap, a prerequisite for frequent replica exchanges. The curves depicted in Figure 2 sample a broad energy range, with the histograms being centered at -7420 kJ mol^{-1} and -6160 kJ mol^{-1} respectively for the unmodified and the most strongly modified force fields. During the 100 ns simulation, each replica sampled all force fields, although some spent most of their time in only a subset of them (Table 2). To give a more complete picture of the efficiency with which replicas have repeatedly used the different simulation conditions, we report in Figure 3 the force-field trajectories of the two limit cases, namely, those of replicas 2 and 3. Replica 3 sampled the various force fields most evenly, whereas replica 2 showed the most skewed distribution, nevertheless keeping a good sampling efficiency through all force fields. The average time for a replica to move from the unmodified force field to the most strongly modified one was 1770 ps (188 observations), for the reverse process, 1988 ps (184 observations), while a return to the unmodified force field via the most strongly modified one lasted on average 3886 ps (184 observations).

Table 2. Percentage Sampling of the Individual Force Fields by Each Replica during the H-REMD Simulation[a]

scaling factor	1.000	0.965	0.931	0.898	0.867	0.837	0.808	0.780
replica 1	19.82	18.25	15.43	12.88	12.07	9.60	6.87	5.08
replica 2	4.86	6.12	8.23	9.67	10.89	13.05	19.72	27.46
replica 3	10.74	12.19	12.98	12.23	13.65	13.34	13.57	11.30
replica 4	7.88	10.35	12.79	14.98	16.25	15.10	12.17	10.48
replica 5	6.60	7.59	9.56	13.58	14.66	15.77	15.95	16.29
replica 6	20.81	17.45	14.15	11.09	9.05	9.15	8.65	9.65
replica 7	6.94	8.65	10.40	12.78	13.75	15.21	16.24	16.03
replica 8	22.35	19.40	16.46	12.79	9.68	8.78	6.83	3.71

[a] Percentages refer to the sampling of the individual force fields by each replica for the whole simulation period of 100 ns of the H-REMD simulation. The individual force fields are represented by the scaling factors *f* used to modify the electrostatic and LJ parameters (see Section 2.2), whereas replicas are numbered arbitrarily.

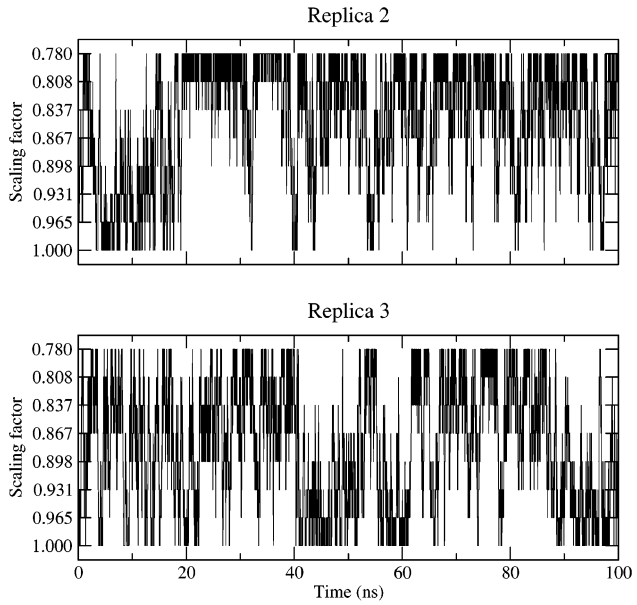


Figure 3. Force-field trajectories of two selected replicas of the H-REMD simulation. Both replicas sample each force field several times. Among all of the replicas, number 2 displays the most skewed and number 3 the most even distribution of the sampled force fields. The individual force fields are indicated on the *y* axis by the scaling factors *f* used to modify the nonbonded interaction parameters (see Section 2.2).

To account for the simulations being performed with different force fields, we assigned to each time point of each replica a statistical weight using WHAM (eqs 11 and 12). When normalized by the average weight assigned to the structures generated using the unmodified force field, the coordinates produced with the most strongly modified force field were assigned an average weight of ca. 5×10^{-9}. Thus, the influence on thermodynamic quantities of the data generated with the most strongly modified force fields is on average very small. However, a certain overlap still exists even between the most extreme situations, since the lowest weight assigned to any frame simulated with the unmodified force field is smaller than the highest weight assigned to any frame simulated with the most strongly modified Hamiltonian. So it is not the case that one could simply omit the data produced by the strongly modified force fields.

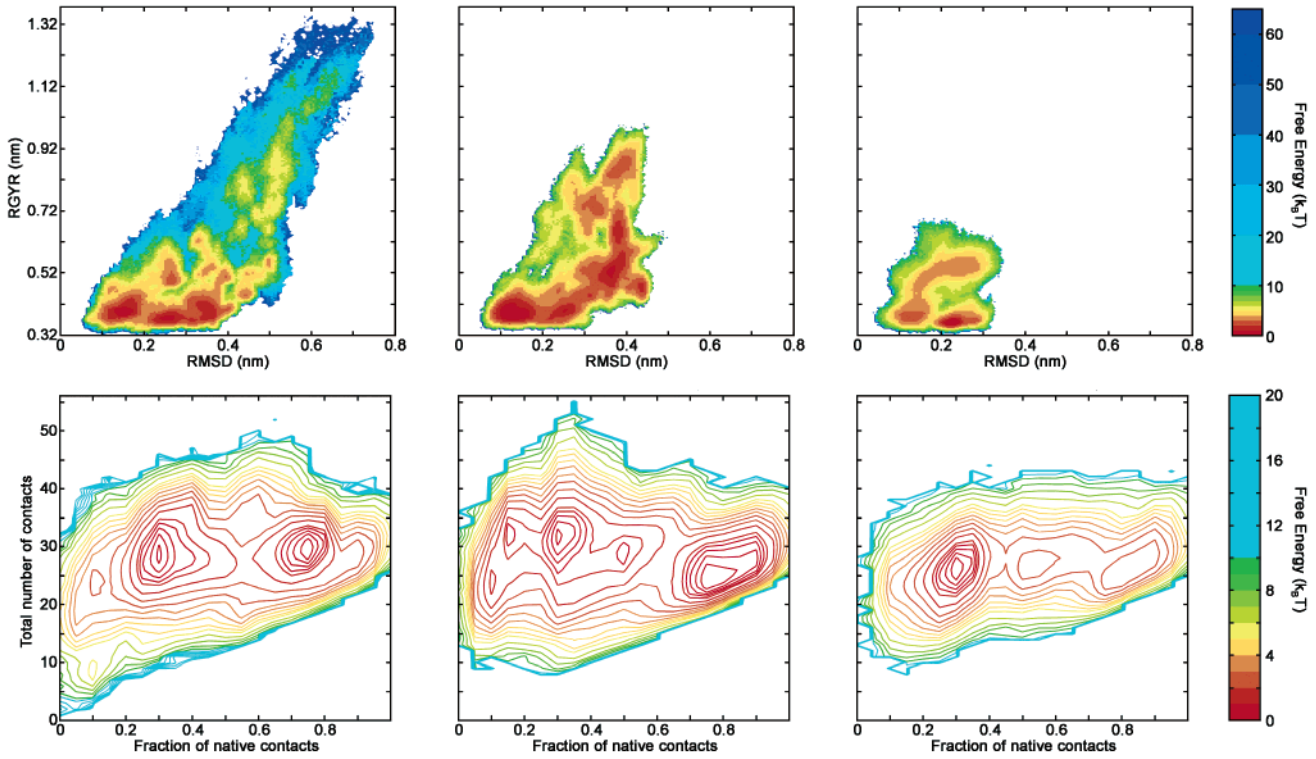


Figure 4. Comparison of two-dimensional representations of the free-energy landscapes for the three simulation approaches we have used. The upper row shows representations resting on the RMSD of the backbone atoms of residues 3−16 from the reference structure and the radius of gyration of the hydrophobic-core atoms (for structural details, see Figure 1). The lower row shows the projection of the free-energy landscapes on the plane defined by the FNC and the total number of contacts. The first column refers to the WHAM-weighted results of the H-REMD simulation, the second column to those of the 8CMD simulation, and the third to those of the LCMD simulation. The first 25 ns of each simulation were omitted from the analysis. Contour lines are drawn every 0.2 k_BT between 0 and 1 k_BT, every 0.5 k_BT in the interval 1−5 k_BT, every 1 k_BT between 5 and 10 k_BT, and every 2 k_BT in the interval 10−20 k_BT. All graphs are normalized to a minimum of 0 k_BT. Figures were generated with MATLAB 6.5.

3.2. Free-Energy Landscapes. Enhanced sampling of the conformational space entails avoiding a trapping of the system within local energy minima through repeated exchanges of the simulation conditions among replicas. Therefore, a way to judge the efficiency in conformational space sampling is the comparison of two-dimensional representations of the FEL computed from structural ensembles simulated with different methods. After having tested several global parameters for the FEL representations, we have selected those that most clearly describe the situation, namely, the backbone RMSD of residues 3−16 relative to our reference structure (for details, see Figure 1), the radius of gyration of the hydrophobic core (RGYR), the FNC, and the total number of contacts. The FEL representations thus obtained from our three simulation approaches are depicted in Figure 4. By far, the LCMD simulation samples the smallest region of the conformational space (third column of Figure 4). For the 8CMD simulations, the volume of the visited conformational space is considerably increased (middle column in Figure 4), but with our H-REMD approach, we were able to increase it even further. The states sampled during the H-REMD simulation reach RGYR values 3.5-fold higher than that of the reference structure, equal to 0.374 nm, and a RMSD greater than 0.7 nm, compared respectively to a 2.8-fold-increased RGYR and a largest RMSD value of ∼0.5 nm observed in the 8CMD simulation. A careful comparison of the two left-most FEL representations in the top row of Figure 4 reveals that minima characterized by increasing RGYR and RMSD values are less well-defined in the H-REMD compared to the 8CMD simulation. We can envisage several plausible explanations for this phenomenon. The first is intrinsically related to the REMD approach, which does not allow the system to extensively explore energy minima. Furthermore, low probability states, likely to be characterized by high values of RGYR and RMSD, are given full weight in the computation of the FEL for the 8CMD simulations, but not so in the case of the H-REMD simulation due to WHAM weighting. Finally, as discussed in Section 3.3, neither simulation approach has reached structural convergence; therefore, the corresponding FEL representations cannot be expected to match completely.

From the FELs shown in Figure 4, it appears that CRT18 frequently visits two states during the simulations, one quite close to the reference structure and the other clearly different, although still characterized by a compact hydrophobic core, but with a small FNC. For a computation of relative free energies to be statistically relevant, it is necessary that a system repeatedly move back and forth between individual states. A projection of the trajectories onto the two-dimensional FELs representations offers a simple way to test if this condition is satisfied. For instance, in the FELs presented in the second row of Figure 4, we can define the

region characterized by a FNC below 0.3 as unfolded, and above 0.75 as folded, and thereafter analyze how often the system moves back and forth among the two regions. Using snapshots taken at 0.2 ps intervals, we observe a total of 30 folding and 32 unfolding events in the H-REMD simulation between 25 and 100 ns, compared to 7 and 12 in the 8CMD, or 6 and 7 during the 25−450 ns LCMD simulation period, none of which occurring in the first 100 ns. The choice of other boundaries for the definition of folded and unfolded regions does not significantly influence the results. For example, defining the boundaries at FNC values of 0.35 and 0.7 yields 102 folding and 105 unfolding events in the H-REMD simulation, compared to 51 and 55 in the 8CMD, or 27 and 28 for the period 25−450 ns of the LCMD simulation, again none of which occurring in the first 100 ns. These results prove that transitions between individual free-energy minima occur more easily in the H-REMD simulation than in either of the CMD ones, thus, implying a better definition of the minima in the FEL representations corresponding to the H-REMD (left column of Figure 4) compared to the ones in the 8CMD (central column) and LCMD (right column) simulations.

In our simulations, the FNC ranges from 1 to 0. The complete sampling of this structural parameter allows us to analyze the percentage of formation of single contacts as a function of the FNC, with the aim of identifying contacts possibly involved in early folding events. For this purpose, we partitioned the FNC into evenly spaced bins, and each snapshot of the simulated ensembles was assigned to the bin corresponding to its FNC. For each bin, we then calculated the fraction of times any given contact was present (data not shown). Although one cannot expect to identify true folding paths from a REMD simulation, the results of this analysis indicate that folding of CRT18 might be facilitated by the formation of contacts between residues 8 and 12, at the tip of the hairpin, as well as between tryptophans 7 and 15.

3.3 Clustering. As an alternative means to quantify conformational space sampling, one can perform a structural clustering and compare both the total number of clusters obtained from the entire coordinates ensemble and how the number of clusters changes as a function of time. We carried out a structural clustering based on a matrix of pairwise RMSDs, following the algorithm described in Section 2.6, which guarantees the distance between cluster centers to be not smaller than the chosen cutoff value. Generally, using snapshots of simulations taken every 10 ps is considered sufficient, since large conformational changes are not expected to occur on this time scale.[43] Instead of taking snapshots, we have used, for our analyses, coordinate averages over 5 ps such that, when dealing with REMD simulation data, cluster weights could be computed from the WHAM weights of their members. Figure 5 shows the results of this analysis for the H-REMD and the 8CMD simulations. Using a cutoff of 0.1 nm for the RMSD between the backbone atoms of residues 3−16, we identified 798 clusters in the H-REMD simulation and only 273 in the 8CMD simulations. Comparing the number of clusters with a weight larger than a certain threshold (panels B−D in Figure 5)

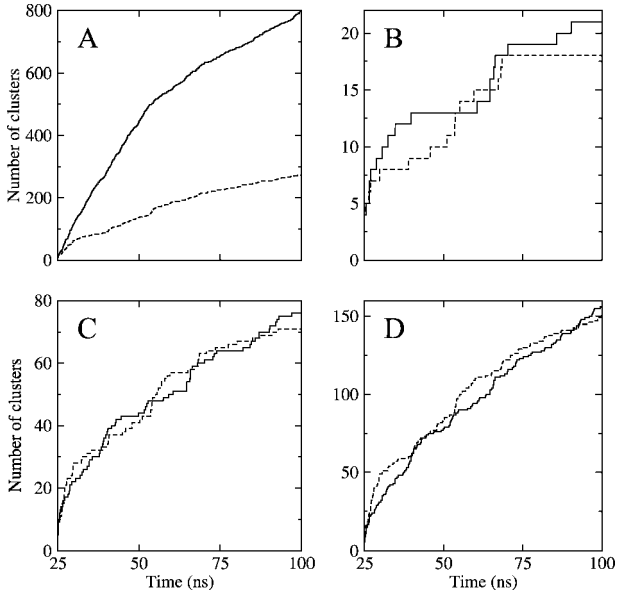


Figure 5. Number of clusters sampled as a function of time during the H-REMD (solid lines) and the 8CMD (dashed lines) simulations, as computed from the whole ensemble of conformations produced by each simulation. Panel A refers to the time courses of the total number of clusters, panel B to those of clusters with weights greater than 1% of all conformations considered, panel C to the ones with weights larger than 0.1%, and panel D to clusters whose weight exceeds 0.01%. The analysis has been done on coordinate averages over 5 ps to allow, in the case of H-REMD, a weighting of the clusters based on the WHAM weights of their members. For additional details, see Section 2.6.

reveals that this difference mainly stems from poorly populated clusters with weights below 0.01%. The final slopes of the curves depicted in Figure 5 show that, with neither simulation approach, clustering has reached convergence. However, comparing the time courses of the total number of clusters and those with weights greater than 1%, namely, neglecting structures simulated with strongly modified force fields, further confirms the improvement in conformational space sampling obtained with our H-REMD approach compared to 8CMD. For instance, after 40 ns, 13 clusters with final weights greater than 1% are already defined in the H-REMD data set, compared to 9 in the 8CMD simulations. To analyze the H-REMD simulation, we also tested a clustering algorithm slightly different from the one described in the Methods and Analysis section, in which we defined the center of a cluster to be the structure whose neighbors have the largest sum of WHAM weights, rather than that with the largest number of neighbors. The two algorithms produce comparable results, the structures of the largest cluster centers being very similar, as well as the cluster sizes (data not shown). The only notable differences are the total number of clusters (822 as compared to 798) and an increase in the number of clusters formed by only a few members. On the basis of these findings, we decided to use the clustering algorithm described in Section 2.6, also because it can be applied to both the H-REMD and the 8CMD simulations, thus allowing a direct comparison of the results.

Figure 5 shows that the H-REMD and the 8CMD simulations display a comparable time evolution of the number of clusters with weights greater than 0.1% (panel C) and 0.01% (panel D). Since there is an apparently large overlap between the FEL representations computed from the H-REMD and those that stem from the 8CMD simulations (Figure 4), one could argue that the two simulations sample the same region of the conformational space and that the H-REMD approach simply extends the conformational sampling toward unfolded structures. However, comparing the conformations corresponding to the centers of the 20 most populated clusters reveals that many of those occurring in one simulation are not sampled by the other. This difference can easily be rationalized considering that our representations of the FEL are highly degenerate. Hence, basins that appear to be well-defined actually result from the superposition of numerous minima, which cannot be distinguished by the collective entities used to reduce a hyperdimensional space to two dimensions. This explains the apparent large overlap between the two-dimensional FEL representations computed from the H-REMD and the 8CMD simulations, despite the relative dissimilarity between the 20 most populated cluster centers. Still, as can be seen in Figure 6, the structures corresponding to the center of the largest cluster of the H-REMD (structure B) and 8CMD (structure C) simulations are almost identical (backbone RMSD of 0.076 nm, heavy-atoms RMSD of 0.168 nm) and, if one neglects the N- and C-terminal residues, highly similar to the reference structure. Furthermore, the largest clusters are of comparable size in the two simulations, being populated by 15.68% and 18.60% respectively for the H-REMD and 8CMD approaches. A totally different situation applies to the LCMD simulation, where the most populated cluster (structure D in Figure 6) is clearly different from the reference structure.

Figure 6 further illustrates the improved efficiency of our H-REMD protocol, compared to 8CMD and LCMD, in sampling the conformational space. The configurations corresponding to the center of the cluster with the largest RMSD from the reference structure are still quite compact in the 8CMD and LCMD simulations (respectively structures F and G in Figure 6), whereas that relative to the H-REMD simulation with an unmodified force field (structure E in Figure 6) is almost fully extended. A visual inspection of the atomic trajectories of the single replicas in the H-REMD simulation reveals that, during the simulation, CRT18 completely unfolded—and refolded to a compact state—in several replicas.

3.4. Correlation Coefficients. The results discussed so far consistently show that our H-REMD protocol is more efficient in sampling the conformational space compared to classical simulation approaches. The clustering analysis just discussed also provides some information in this respect. As another independent approach to compare the simulation protocols used in our investigation, we have computed with eq 14 the correlation coefficients depicted in Figure 7. We used 100 bins of sizes 0.01 nm and 0.01 respectively for RMSD (left panel) and FNC (right panel), whereas t was incrementally increased by 100 ps, starting from 25 ns up to 100 ns, leading to 750 P distributions. The time courses

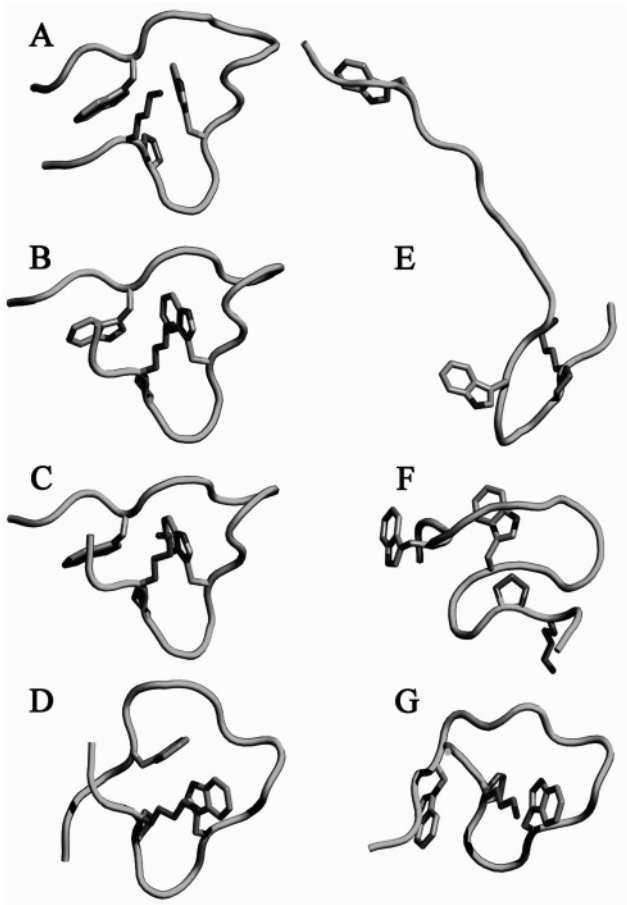


Figure 6. Models of the reference structure of CRT18 (A), of the conformations corresponding to the centers of the most populated clusters (B for H-REMD, C for 8CMD, and D for LCMD), and of the centers of the clusters with the largest RMSD from structure A (E for H-REMD with $f = 1$, F for 8CMD, and G for LCMD). Structures averaged over 5 ps intervals were used for the clustering, excluding the initial 25 ns of each simulation, which resulted in a total of 1.2×10^5 structures for both the H-REMD and the 8CMD simulations and 8.5×10^4 structures for the LCMD simulation. The backbone/all-atoms RMSD from A, for residues 3−16, are 0.1045/0.2042 nm for B, 0.0983/0.2234 nm for C, and 0.2240/0.4276 nm for D. The models were drawn using PYMOL (http://www.pymol.org).

of the correlation coefficient for both parameters show that the H-REMD simulation (continuous curves) converges more rapidly compared to the 8CMD (dashed traces) simulation, the difference being particularly pronounced for the FNC.

3.5. Comparison with NMR Data. During a simulation, one not only wishes to sample as large a volume of the conformational space as possible but also reproduce structural features measured experimentally. We have, therefore, compared our simulated ensembles with the upper bounds of inter-residue interproton distances d derived from NOE measurements at 278 K. The whole set of 126 unambiguous interproton distances identified from NMR data (see Section 2.8) has been used for the comparison. The graphs in the top-left, top-right, and bottom-left panels of Figure 8 show the ensemble averages $\langle d^{-6} \rangle^{-1/6}$, computed respectively from the 8CMD, H-REMD, and LCMD simulations, plotted

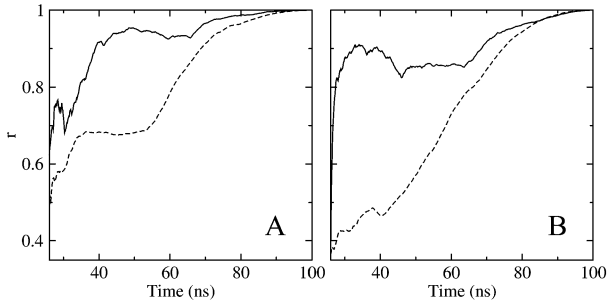


Figure 7. Time series of the correlation coefficients computed with eq 14 for the RMSD of the backbone atoms of residues 3−16, relative to the reference structure (left panel), and for the FNC (right panel). The solid lines refer to WHAM-weighted H-REMD data, whereas the dashed curves correspond to the 8CMD simulations. More details are given in Section 2.7.

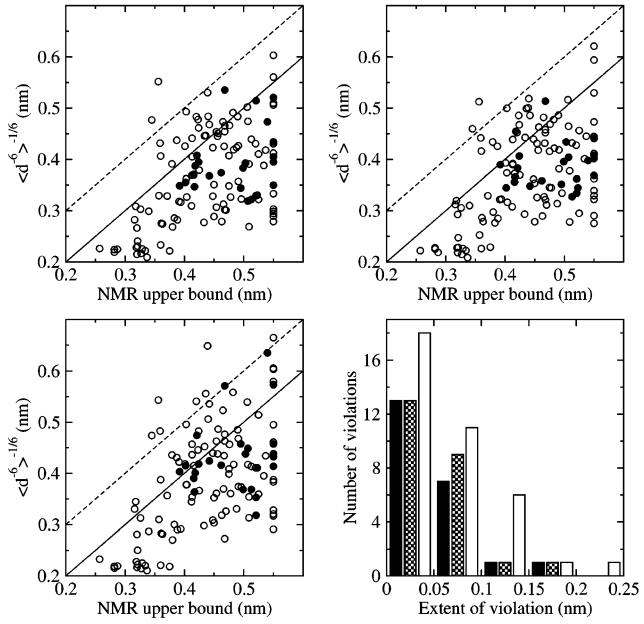


Figure 8. Simulated ensemble averages of interproton distances $\langle d^{-6}\rangle^{-1/6}$, plotted against upper bounds of the same distances derived from NOE measurements. A set of 126 unambiguous interproton distances was used for the analysis (see Section 2.8). The upper-left panel refers to the simulation 8CMD, the upper-right to the H-REMD, and the lower left to the LCMD. In all three cases, filled circles represent distances between protons belonging to amino acids at least four residues apart along the polypeptide chain. The solid diagonal lines indicate the boundary above which the experimental distances are violated, whereas the dashed lines represent the boundary above which such distances are violated by more than 0.1 nm. In the lower-right panel, the number of violations is plotted against their extent, with full, dashed, and empty bars referring respectively to the 8CMD, H-REMD, and LCMD simulations. In the case of the H-REMD data, ensemble averages were calculated using WHAM weights.

against the upper bounds of the same interproton distances derived from NMR data. Distinction is made for distances involving amino acids at least four residues apart along the polypeptide chain (filled circles) because they are most indicative of the global structure of the molecule. To give an overview of the situation, the bottom-right panel of Figure 8 displays the number of violations relative to four distance ranges. The LCMD simulation displays the largest number of violations, in total 37. Instead, despite the diversity of their conformational ensembles, the violations for the H-REMD and 8CMD simulations are comparably few (24 for the former and 22 for the latter), also if only the structures simulated with the unmodified force field are included in the analysis (data not shown). The situation does not change if one considers only the most significant violations, namely, those greater than 0.05 nm (19 for LCMD, 11 for H-REMD, and 9 for 8CMD). Combining the sets of such interproton distance violations occurring in either the H-REMD or 8CMD simulations yields a list of 13 proton pairs, of which 9 involve TRP-7, as shown in Table 3. Clearly, the strongest violations observed in the two approaches involve hydrogen pairs belonging to PRO-4 and TRP-7.

Table 3. Violated Distances between Individual Pairs of Hydrogen Atoms[a]

atoms involved		violations (nm)	
atom 1/residue	atom 2/residue	H-REMD	8CMD
HB2/PRO-4	HN/ASP-6	0.056	0.046
HB2/PRO-4	HN/TRP-7	0.156	0.196
HB2/PRO-4	HB1/TRP-7	0.114	0.131
HB2/PRO-4	HE3/TRP-7	0.032	0.063
HG2/PRO-4	HN/TRP-7	0.070	0.053
HG2/PRO-4	HB1/TRP-7	0.063	0.054
HG2/PRO-4	HE3/TRP-7	0.083	0.072
HD1/TRP-7	HB2/GLU-9	0.061	0.091
HZ2/TRP-7	HB1/ASP-12	0.045	0.067
HB1/ASP-8	HB2/MET-11	0.055	0.021
HB1/ASP-8	HG1/MET-11	0.078	0.044
HB1/MET-11	HA/ASP-12	0.096	0.090
HE1/TRP-15	HD1/PRO-18	0.063	0.010

[a] The list includes all hydrogen pairs involved in interproton inter-residue distances violated by more than 0.05 nm with respect to the NMR data in either the H-REMD or the 8CMD simulation.

While discussing the time evolution of the FNC, we have postulated the formation of contacts between residues 8 and 12 and between tryptophans 7 and 15 to be early events during folding (next-to-last paragraph of the subsection on free-energy landscapes). Interestingly, the comparison of the simulated ensembles with NMR data shows that the four unequivocal interproton, inter-residue distances involving tryptophans 7 and 15 are never violated during the LCMD and 8CMD simulations, and only one of them, namely, Trp7(HZ2)−Trp15(HA), is moderately violated (by 0.011 nm) exclusively in the LCMD simulation. Furthermore, in all three structural ensembles, the only Asp8−Asp12 unequivocal distance derived from NMR data (HB1−HN) is again never violated.

We have analyzed the ensemble averages $\langle d^{-6}\rangle^{-1/6}$ separately for the individual clusters (data not shown) and found that each of them violates more of the NMR-derived interproton distances, and more severely than if the ensemble average is taken for the whole simulation. This applies even to the most populated cluster, which is very similar to the reference structure. None of the found clusters fits the experimental data better than the whole ensemble of simu-

lated structures. This finding is in agreement with previous reports[45] indicating that ensemble averages more accurately reproduced NMR data than single structures.

4. Conclusions

Judging from the simulations on the CRT18 fragment reported here, our novel H-REMD approach considerably enhances conformational space sampling compared to classical MD, without affecting the computational effort. The better sampling also involves structures with very small Boltzmann weights, which, however, can potentially have important biological functions. No direct comparison with previous reports on H-REMD simulations on biological macromolecules can be done because the simulation conditions are too different, as described at the beginning of the Results and Discussion section. Compared to T-REMD, our approach allows a fine-tuning of the different simulation conditions used for REMD to specifically influence those degrees of freedom that are of greatest interest, or that are believed to most strongly affect the dynamic properties of the system being investigated. This allows a substantial reduction in the number of replicas needed to perform a REMD simulation, while still permitting an efficient conformational searching. Our results show that the data produced by all the different force fields used for a H-REMD simulation can be successfully combined with WHAM. Furthermore, simulating with modified force fields does not introduce structural artifacts, as demonstrated by the comparison with NMR data and with classical simulation approaches. The results reported here support our premise that acting on the frustration of nonbonded interactions influences conformational-space sampling. Last, but not least, our simulations favor the hypothesis that, contrary to what has been observed for longer fragments of the calreticulin P domain, CRT18 does not fold to a unique structure.

Acknowledgment. This work was supported by the ETH Zurich with Grants 0-20915-01 and 0-50590-04. The authors thank Pascal Bettendorff and Lars Ellgaard for kindly providing NMR data on CRT18 prior to their publication and Daniele Passerone, Lars Ellgaard, and Xavier Daura for stimulating discussions. The constructive criticisms and suggestions by the reviewers of this article are also kindly acknowledged.

References

(1) Balabin, L. A.; Onuchic, J. N. Dynamically Controlled Protein Tunneling Paths in Photosynthetic Reaction Centers. *Science* **2000**, *290*, 114−117.

(2) Mitsutake, A.; Sugita, Y.; Okamoto, Y. Generalized-ensemble algorithms for molecular simulations of biopolymers. *Biopolymers* **2001**, *60* (2), 96−123.

(3) Berg, B. A.; Neuhaus, T. Multicanonical algorithms for first-order phase transitions. *Phys. Lett.* **1991**, *267* (2), 249−53.

(4) Berg, B. A.; Neuhaus, T. Multicanonical ensemble: A new approach to simulate first-order phase transitions. *Phys. Rev. Lett.* **1992**, *68* (1), 9−12.

(5) Lyubartsev, A. P.; Martinovski, A. A.; Shevkunov, S. V.; Vorontsov-Velyaminov, P. N. New approach to Monte Carlo calculations of the free energy: Method of expanded ensembles. *J. Chem. Phys.* **1992**, *96* (3), 1776−83.

(6) Marinari, E.; Parisi, G. Simulated tempering: A new Monte Carlo scheme. *Europhys. Lett.* **1992**, *19*, 451−58.

(7) Sugita, Y.; Okamoto, Y. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* **1999**, *314*, 141−51.

(8) Sugita, Y.; Okamoto, Y. Replica-exchange multicanonical algorithm and multicanonical replica-exchange method for simulating systems with rough energy landscape. *Chem. Phys. Lett.* **2000**, *329*, 261−70.

(9) Sugita, Y.; Kitao, A.; Okamoto, Y. Multidimensional replica-exchange method for free-energy calculations. *J. Chem. Phys.* **2000**, *113* (15), 6042−51.

(10) Yamamoto, R.; Kob, W. Replica-exchange molecular dynamics simulation for supercooled liquids. *Phys. Rev. E: Stat. Phys., Plasmas, Fluids, Relat. Interdiscip. Top.* **2000**, *61* (5B), 5473−6.

(11) Okabe, T.; Kawata, M.; Okamoto, Y.; Mikami, M. Replica-exchange Monte Carlo method for the isobaric−isothermal ensemble. *Chem. Phys. Lett.* **2001**, *335*, 435−9.

(12) Sanbonmatsu, K. Y.; Garcia, A. E. Structure of Met-enkephalin in explicit aqueous solution using replica exchange molecular dynamics. *Proteins* **2002**, *46* (2), 225−34.

(13) Hukushima, K. Domain-wall free energy of spin-glass models: numerical model and boundary conditions. *Phys. Rev. E.: Stat., Nonlinear, Soft Matter Phys.* **1999**, *60* (4), 3606−3614.

(14) Gront, D.; Kolinski, A.; Skolnick, J. Comparison of three Monte Carlo conformational search strategies for a protein-like homopolymer model: Folding thermodynamics and identification of low-energy structures. *J. Chem. Phys.* **2000**, *113* (12), 5065−71.

(15) Hukushima, K.; Nemoto, K. Exchange Monte Carlo method and application to spin glass simulations. *J. Phys. Soc. Jpn.* **1996**, *65* (4), 1604−8.

(16) Fukunishi, H.; Watanabe, O.; Takada, S. On the Hamiltonian replica exchange method for efficient sampling of biomolecular systems: Application to protein structure prediction. *J. Chem. Phys.* **2002**, *116* (20), 9058−67.

(17) Murata, K.; Sugita, Y.; Okamoto, Y. Molecular dynamics simulations of DNA dimers based on replica-exchange umbrella sampling II: Free energy analysis. *J. Theor. Comput. Chem.* **2005**, *4*, 433−448.

(18) Snow, C. D.; Sorin, E. J.; Rhee, Y. M.; Pande, V. How well can simulation predict protein folding kinetics and thermodynamics. *Annu. Rev. Biophys. Biomol. Struct.* **2005**, *34*, 43.

(19) Tavernelli, I.; Di Iorio, E. E. The interplay between protein dynamics and frustration of nonbonded interactions as revealed by molecular dynamics simulations. *Chem. Phys. Lett.* **2001**, *345*, 287−294.

(20) Wolynes, P. G.; Eaton, W. A. The physics of protein folding. *Phys. World* **1999**, *12*, 39−44.

(21) Bryngelson, J. D.; Onuchic, J. N.; Socci, N. D.; Wolynes, P. G. Funnels, pathways, and the energy landscape of protein folding: A synthesis. *Proteins* **1995**, *21* (3), 167−95.

(22) Piela, L.; Kostrowicki, J.; Scheraga, H. A. The multiple-minima problem in the conformational analysis of molecules. Deformation of the potential energy hypersurface by the diffusion equation method. *J. Chem. Phys.* **1989**, *93*, 3339−3346.

(23) Kostrowicki, J.; Scheraga, H. A. Application of the diffusion equation method for global optimization to oligopeptides. *J. Chem. Phys.* **1992**, *96*, 7442−7449.

(24) Schug, A.; Herges, T.; Wenzel, W. Reproducible protein folding with the stochastic tunneling method. *Phys. Rev. Lett.* **2003**, *91* (15), 158102.

(25) Pappu, R. V.; Marshall, G. R.; Ponder, J. W. A potential smoothing algorithm accurately predicts transmembrane helix packing. *Nat. Struct. Mol. Biol.* **1999**, *6* (1), 50−55.

(26) Ellgaard, L.; Frickel, E.-M. Calnexin, calreticulin and ERp57: teammates in glycoprotein folding. *Cell Biochem. Biophys.* **2003**, *39*, 223−248.

(27) Ellgaard, L.; Bettendorff, P.; Braun, D.; Herrmann, T.; Fiorito, F.; Jelesarov, I.; Guntert, P.; Helenius, A.; Wuthrich, K. NMR structures of 36 and 73-residue fragments of the calreticulin P-domain. *J. Mol. Biol.* **2002**, *322* (4), 773−84.

(28) Ellgaard, L.; Riek, R.; Braun, D.; Herrmann, T.; Helenius, A.; Wuthrich, K. Three-dimensional structure topology of the calreticulin P-domain based on NMR assignment. *FEBS Lett.* **2001**, *488* (1−2), 69−73.

(29) Ellgaard, L.; Riek, R.; Herrmann, T.; Guntert, P.; Braun, D.; Helenius, A.; Wuthrich, K. NMR structure of the calreticulin P-domain. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98* (6), 3133−8.

(30) Berendsen, H. J. C.; van der Spoel, D.; van Drunen, R. GROMACS: A message-passing parallel molecular dynamics implementation. *Comput. Phys. Commun.* **1995**, *91*, 43−56.

(31) Lindahl, E.; Hess, B.; van der Spoel, D. GROMACS 3.0: A package for molecular simulation and trajectory analysis. *J. Mol. Model.* **2001**, *7*, 306−17.

(32) van Gunsteren, W. F.; Billeter, S. R.; Eising, A. A.; Hünenberger, P. H.; Krüger, P.; Mark, A. E.; Scott, W. R. P.; Tironi, I. G. *Biomolecular simulation: The GROMOS96 manual and user guide*; vdf Hochschulverlag AG an der ETH Zurich, BIOMOS b.v. Zurich Groningen: Zurich, Switzerland; Groningen, Netherlands, 1996.

(33) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Di Nola, A.; Haak, J. R. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **1984**, *81*, 3684−3690.

(34) Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M. LINCS: A linear constraint colver for molecular simulations. *J. Comput. Chem.* **1997**, *18*, 1463−72.

(35) Miyamoto, S.; Kollman, P. A. SETTLE: An analytical version of the SHAKE and RATTLE algorithms for rigid water models. *J. Comput. Chem.* **1992**, *13*, 952−62.

(36) Tironi, I. G.; Sperb, R.; Smith, P. E.; van Gunsteren, W. F. A generalized reaction field method for molecular dynamics simulations. *J. Chem. Phys.* **1995**, *102*, 5451−5459.

(37) Smith, P. E.; van Gunsteren, W. F. Consistent dielectric properties of the simple point charge and extended simple point charge water models at 277 and 300 K. *J. Chem. Phys.* **1994**, *100* (4), 3169−74.

(38) Berendsen, H. J. C.; Grigera, J. R.; Straatsma, T. P. The missing term in effective pair potentials. *J. Chem. Phys.* **1987**, *91* (24), 6269−71.

(39) Kumar, S.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A.; Rosenberg, J. M. The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *J. Comput. Chem.* **1992**, *13* (8), 1011−21.

(40) Kumar, S.; Rosenberg, J. M.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A. Multidimensional free-energy calculations using the weighted histogram analysis method. *J. Comput. Chem.* **1995**, *16* (11), 1339−50.

(41) Boczko, E. M.; Brooks, C. L., III. Constant-temperature free energy surfaces for physical and chemical processes. *J. Phys. Chem.* **1993**, *97* (17), 4509−13.

(42) Tavernelli, I.; Cotesta, S.; Di Iorio, E. E. Protein dynamics, thermal stability, and free-energy landscapes: a molecular dynamics investigation. *Biophys. J.* **2003**, *85* (4), 2641−9.

(43) Daura, X.; Gademann, K.; Jaun, B.; Seebach, D.; van Gunsteren, W. F.; Mark, A. E. Peptide folding: When simulation meets experiment. *Angew. Chem., Int. Ed.* **1999**, *38*, 236−40.

(44) Jang, S.; Shin, S.; Pak, Y. Replica-exchange method using the generalized effective potential. *Phys. Rev. Lett.* **2003**, *91* (5), 058305.

(45) Daura, X.; Gademann, K.; Schafer, H.; Jaun, B.; Seebach, D.; van Gunsteren, W. F. The beta-peptide hairpin in solution: conformational study of a beta-hexapeptide in methanol by NMR spectroscopy and MD simulation. *J. Am. Chem. Soc.* **2001**, *123* (10), 2393−404.

(46) Kabsch, W.; Sander, C. Dictionary of secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **1983**, *22*, 2577−2637.

CT050250B

JCTC Journal of Chemical Theory and Computation

Detailed Balance in Ehrenfest Mixed Quantum-Classical Dynamics

Priya V. Parandekar and John C. Tully*

Department of Chemistry, Yale University, P.O. Box 208107, New Haven, Connecticut 06520

Received August 25, 2005

Abstract: We examine the equilibrium limits of self-consistent field (Ehrenfest) mixed quantum-classical dynamics. We derive an analytical expression for the equilibrium mean energy of a multistate quantum oscillator coupled to a classical bath. We show that, at long times, for an ergodic system, the mean energy of the quantum subsystem always exceeds the temperature of the classical bath that drives it. Furthermore, the energy becomes larger as the number of states increases and diverges as the number of quantum levels approaches infinity. We verify these results by simulations.

1. Introduction

Mixed quantum-classical dynamics (MQCD), in which selected quantum mechanical degrees of freedom are coupled to a system of classical mechanical degrees of freedom, has proved to be a useful complement to standard classical molecular dynamics (MD) simulations. Quantum effects including electronic transitions,[1,2] proton tunneling, and zero-point motion[3−6] can be introduced within a computationally tractable classical MD framework. A critical requirement for the success of a MQCD theory is the proper treatment of the "quantum backreaction", the altering of the classical forces due to transitions in the quantum subsystem.[7−12] Two widely used approaches for approximating the quantum backreaction have emerged, "surface hopping"[13−15] and "Ehrenfest".[16−20] In both approaches, quantum transitions arise in the same way, governed by the time-dependent Schrödinger equation in which the time variation of the Hamiltonian arises from the motions of the classical particles. The methods differ only in the way the classical paths evolve. In surface hopping, the forces derive from a single quantum state, subject to sudden stochastic "hops" to different quantum states. The Ehrenfest method is a self-consistent field method; the forces governing the classical particles arise from a weighted average of quantum states. Both the surface-hopping and Ehrenfest methods allow for energy transfer between the quantum and classical subsystems such that the total energy is conserved, and both methods have proved quite accurate in many applications.

We showed in a previous paper that, for a two-level quantum system coupled to a many-particle classical bath, the "fewest-switches" version of surface hopping[13] correctly obeys detailed balancing; the two-level system approaches a quantum temperature equal to the classical temperature of the bath.[21] This is not necessarily the case for the Ehrenfest method, as has been discussed by several authors,[21−24] signaling a potentially serious deficiency of the method. We previously derived a closed-form expression for the mean energy of a two-level quantum system coupled by Ehrenfest dynamics to a classical bath, showing that the quantum subsystem approaches a temperature that is finite but higher than the temperature of the classical bath to which it is coupled.[21] In this paper, we generalize our previous result to a quantum subsystem composed of an arbitrary number of quantum states. For the special case of equally spaced quantum levels, we are able to obtain an exact, closed-form expression for the mean energy of the quantum subsystem. This expression shows, remarkably, that in the limit of an infinite number of equally spaced levels, that is, a harmonic oscillator, the mean energy of the quantum subsystem approaches infinity no matter how low the classical temperature that drives it. The populations of each state are not equal in this limit, so the quantum subsystem does not approach infinite temperature. However, the populations decrease with increasing quantum number sufficiently slowly so that the mean energy diverges.

* Corresponding author phone: (203) 432−3934; fax: (203) 432−6144; e-mail: john.tully@yale.edu.

10.1021/ct050213k CCC: $33.50
Published on Web 01/10/2006

2. Two-Level Quantum Subsystem

In a previous publication,[21] we derived a closed-form expression for the equilibrium mean energy of a two-level quantum subsystem coupled to an infinite number of classical particles via the Ehrenfest self-consistent field approximation. We recast the amplitudes c_α and c_β of quantum levels α and β into two new variables

$$X = |c_\beta|^2 = 1 - |c_\alpha|^2 \tag{1a}$$

and

$$Y = c_\alpha^* c_\beta + c_\beta^* c_\alpha \tag{1b}$$

The variables X and Y can be shown to behave as effective classical variables. [25] We derived a classical Liouville equation for the probability distribution $f(\mathbf{q},\mathbf{p},X,Y)$ of the positions $\mathbf{q}$ and momenta $\mathbf{p}$ of the classical particles and the variables X and Y. We then obtained the steady-state solution of the Liouville equation, which produced the following simple expression for the equilibrium mean energy of the quantum subsystem in terms of the temperature, T, of the classical bath:

$$\bar{E} = \langle |c_\beta|^2 \rangle \epsilon_\beta = k_B T - \frac{\epsilon_\beta \exp[-\epsilon_\beta/k_B T]}{1 - \exp[-\epsilon_\beta/k_B T]} \tag{2}$$

where the energy of the lower quantum level α is taken to be zero, ϵ_β is the energy of the upper level β, and k_B is Boltzmann's constant. It can be shown with some algebra that the mean energy of the quantum subsystem produced by Ehrenfest dynamics, as given by eq 2, is always greater than the desired Boltzmann energy

$$\bar{E}_{BOLTZ} = \frac{\epsilon_\beta \exp[-\epsilon_\beta/k_B T]}{1 + \exp[-\epsilon_\beta/k_B T]} \tag{3}$$

3. Three-Level Quantum Subsystem

We first generalize this result to a three-level quantum subsystem and, later, to an N-level system. We follow the same procedure as above. The three-level quantum subsystem is described in an adiabatic basis. The wave functions α_q, β_q, and γ_q are the eigenfunctions of the quantum Hamiltonian for fixed classical positions $\mathbf{q}$

$$\mathscr{H}_q \alpha_q = \epsilon_\alpha(\mathbf{q})\, \alpha_q \tag{4a}$$

$$\mathscr{H}_q \beta_q = \epsilon_\beta(\mathbf{q})\, \beta_q \tag{4b}$$

$$\mathscr{H}_q \gamma_q = \epsilon_\gamma(\mathbf{q})\, \gamma_q \tag{4c}$$

The subscript q indicates that the quantum Hamiltonian and its eigenfunctions depend parametrically on the classical positions $\mathbf{q}$. The eigenenergies $\epsilon_\alpha(\mathbf{q})$, $\epsilon_\beta(\mathbf{q})$, and $\epsilon_\gamma(\mathbf{q})$, in general, are also functions of $\mathbf{q}$; they are the adiabatic potential energy surfaces for states α, β, and γ, respectively. We express the wave function of the quantum subsystem at any time t as a linear combination of α_q, β_q, and γ_q

$$\psi(t) = c_\alpha(t)\, \alpha_q + c_\beta(t)\, \beta_q + c_\gamma(t)\, \gamma_q \tag{5}$$

where $c_\alpha(t)$, $c_\beta(t)$, and $c_\gamma(t)$ are the complex-valued expansion coefficients. Substituting eq 5 into the time-dependent Schrödinger equation gives a set of coupled differential equations for the time-varying amplitudes $c_\alpha(t)$, $c_\beta(t)$, and $c_\gamma(t)$:

$$\dot{c}_\alpha = -\frac{i}{\hbar}\epsilon_\alpha c_\alpha - \dot{\mathbf{q}}\cdot\mathbf{d}_{\alpha\beta} c_\beta + \dot{\mathbf{q}}\cdot\mathbf{d}_{\gamma\alpha} c_\gamma \tag{6a}$$

$$\dot{c}_\beta = -\frac{i}{\hbar}\epsilon_\beta c_\beta + \dot{\mathbf{q}}\cdot\mathbf{d}_{\alpha\beta} c_\alpha - \dot{\mathbf{q}}\cdot\mathbf{d}_{\beta\gamma} c_\gamma \tag{6b}$$

$$\dot{c}_\gamma = -\frac{i}{\hbar}\epsilon_\gamma c_\gamma - \dot{\mathbf{q}}\cdot\mathbf{d}_{\gamma\alpha} c_\alpha + \dot{\mathbf{q}}\cdot\mathbf{d}_{\beta\gamma} c_\beta \tag{6c}$$

where we have assumed, for simplicity of notation, that the adiabatic wave functions α_q, β_q, and γ_q are real-valued. The nonadiabatic coupling vector $\mathbf{d}_{\alpha\beta}$ is given by

$$\mathbf{d}_{\alpha\beta} = \langle \alpha_q | \nabla_q \beta_q \rangle \tag{7}$$

Couplings between the other states are defined similarly. Note that, because we have chosen the adiabatic representation, there is no potential energy term coupling the quantum levels. Since the Ehrenfest method is invariant to choice of representation,[26] the results we derive here apply for any valid representation. From eq 6, we obtain the time derivatives of the populations.

$$\frac{d}{dt}|c_\alpha|^2 = -\dot{\mathbf{q}}\cdot\mathbf{d}_{\alpha\beta}(c_\alpha^* c_\beta + c_\beta^* c_\alpha) + \dot{\mathbf{q}}\cdot\mathbf{d}_{\gamma\alpha}(c_\alpha^* c_\gamma + c_\gamma^* c_\alpha) \tag{8a}$$

$$\frac{d}{dt}|c_\beta|^2 = \dot{\mathbf{q}}\cdot\mathbf{d}_{\alpha\beta}(c_\alpha^* c_\beta + c_\beta^* c_\alpha) - \dot{\mathbf{q}}\cdot\mathbf{d}_{\beta\gamma}(c_\beta^* c_\gamma + c_\gamma^* c_\beta) \tag{8b}$$

$$\frac{d}{dt}|c_\gamma|^2 = \dot{\mathbf{q}}\cdot\mathbf{d}_{\beta\gamma}(c_\beta^* c_\gamma + c_\gamma^* c_\beta) - \dot{\mathbf{q}}\cdot\mathbf{d}_{\gamma\alpha}(c_\alpha^* c_\gamma + c_\gamma^* c_\alpha) \tag{8c}$$

Summing all parts of eq 8 demonstrates conservation of norm. We define four independent effective classical phase space variables for the quantum subsystem, U, S, W, and X, in terms of the quantum amplitudes:

$$U = |c_\beta|^2 \tag{9}$$

$$S = |c_\gamma|^2 \tag{10}$$

$$W = c_\beta^* c_\gamma + c_\gamma^* c_\beta \tag{11}$$

$$X = c_\beta^* c_\alpha + c_\alpha^* c_\beta \tag{12}$$

The variables U, S, W, and X are the independent variables required to characterize a three-state system. Whereas the three complex-valued amplitudes introduce six variables, two are not independent. The quantity $|c_\alpha|^2$ is determined by conservation of norm, and the variable

$$Y = c_\gamma^* c_\alpha + c_\alpha^* c_\gamma \tag{13}$$

can be expressed as follows in terms of the independent variables:

$$Y = \frac{\pm\sqrt{4US - W^2}\sqrt{4U - 4U^2 - 4US - X^2} + WX}{2U} \tag{14}$$

From eqs 8−12, we obtain equations for the time derivatives of the four effective classical variables U, S, W, and X:

$$\dot{U} = -\dot{\mathbf{q}}\cdot\mathbf{d}_{\beta\gamma}W + \dot{\mathbf{q}}\cdot\mathbf{d}_{\alpha\beta}X \quad (15)$$

$$\dot{S} = \dot{\mathbf{q}}\cdot\mathbf{d}_{\beta\gamma}W - \dot{\mathbf{q}}\cdot\mathbf{d}_{\gamma\alpha}Y \quad (16)$$

$$\dot{W} = 2(U-S)\dot{\mathbf{q}}\cdot\mathbf{d}_{\beta\gamma} - \dot{\mathbf{q}}\cdot\mathbf{d}_{\gamma\alpha}X + \dot{\mathbf{q}}\cdot\mathbf{d}_{\alpha\beta}Y \pm \frac{(\epsilon_\gamma-\epsilon_\beta)}{\hbar}\sqrt{4US-W^2} \quad (17)$$

$$\dot{X} = 2(1-2U-S)\dot{\mathbf{q}}\cdot\mathbf{d}_{\alpha\beta} + \dot{\mathbf{q}}\cdot\mathbf{d}_{\gamma\alpha}W - \dot{\mathbf{q}}\cdot\mathbf{d}_{\beta\gamma}Y \pm \frac{(\epsilon_\alpha-\epsilon_\beta)}{\hbar}\sqrt{4U-4U^2-4US-X^2} \quad (18)$$

The two separate branches (±) in eqs 17 and 18 can be treated separately and, thereby, pose no mathematical difficulty. As the quantum subsystem evolves according to the time-dependent Schrödinger equation, the classical subsystem evolves self-consistently according to Hamilton's equations of motion. As for the two-level case, the phase space variables of the classical subsystem are (**q**, **p**) where **q** and **p** are the positions and momenta, respectively, of the N_c number of classical particles. For simplicity of notation, we assume that the classical variables have been transformed into a frame in which the quantum system is coupled to only a single component of momentum, p_1. As shown elsewhere,[27] using mass-weighted coordinates, this can be achieved for any pair of quantum levels. We assume further that all of the nonadiabatic couplings are in the same direction. The latter is not true in general, but since detailed balance is a statement about the forward and backward transition rates between a pair of quantum states, this simplification cannot affect the final results. As a result of these simplifying assumptions, the dot products can be replaced by scalar multiplications in eqs 15−18. The backreaction of the quantum subsystem on the classical subsystem is incorporated as the Hellmann−Feynman force, which acts only on classical momentum 1.

$$\dot{p}_1 = -\frac{\partial V(\mathbf{q})}{\partial q_1} - \frac{\partial}{\partial q_1}\langle\psi(t)|H_q|\psi(t)\rangle = -\frac{\partial V(\mathbf{q})}{\partial q_1} - (\epsilon_\gamma-\epsilon_\beta)d_{\beta\gamma}W - (\epsilon_\beta-\epsilon_\alpha)d_{\alpha\beta}X - (\epsilon_\alpha-\epsilon_\gamma)d_{\gamma\alpha}Y \quad (19)$$

Now equipped with the time derivatives of the phase space variables (quantum and classical), we proceed to derive the probability distribution function $f(\mathbf{q},\mathbf{p},U,S,W,X)$, which obeys the Liouville equation (see McQuarrie[28])

$$\frac{\partial f}{\partial t} = \sum_{i=1}^{N_c}\left[\frac{p_i}{m}\frac{\partial f}{\partial q_i} - \frac{\partial V}{\partial q_i}\frac{\partial f}{\partial p_i}\right] + \frac{\partial(f\dot{U})}{\partial U} + \frac{\partial(f\dot{S})}{\partial S} + \frac{\partial(f\dot{W})}{\partial W} + \frac{\partial(f\dot{X})}{\partial X} + \frac{\partial f}{\partial p_1}[-(\epsilon_\gamma-\epsilon_\beta)d_{\beta\gamma}W - (\epsilon_\beta-\epsilon_\alpha)d_{\alpha\beta}X - (\epsilon_\alpha-\epsilon_\gamma)d_{\gamma\alpha}Y] = 0 \quad (20)$$

The function $f(\mathbf{q},\mathbf{p},U,S,W,X)$ that satisfies eq 20 is

$$f(\mathbf{q},\mathbf{p},U,S,W,X) = A\,\mathrm{e}^{-V(\mathbf{q})/k_BT}\prod_{i=1}^{N_c}\mathrm{e}^{-p_i^2/2mk_BT}\,g(U,S,W,X) \quad (21)$$

where A is a normalization constant and

$$g(U,S,W,X) = \frac{1}{\pi^2}(4US-W^2)^{-1/2}(4U-4U^2-4US-X^2)^{-1/2}\exp\left[-\frac{U(\epsilon_\beta-\epsilon_\alpha)}{k_BT}\right]\exp\left[-\frac{S(\epsilon_\gamma-\epsilon_\alpha)}{k_BT}\right]\exp\left(-\frac{\epsilon_\alpha}{k_BT}\right) = \frac{1}{\pi^2}(4US-W^2)^{-1/2}(4U-4U^2-4US-X^2)^{-1/2}\exp[-(\sum_{i=\alpha,\beta,\gamma}|c_i|^2\epsilon_i)/k_BT] \quad (22)$$

Equation 22 is the probability distribution function which determines any average properties of the three-state quantum subsystem. Because we have carried out the derivation in the adiabatic representation in which the Hamiltonian is diagonal, the energy of the quantum subsystem depends only on the probabilities $|c_\alpha|^2$, $|c_\beta|^2$, and $|c_\gamma|^2$. We can then integrate eq 22 over dW and dX to obtain the un-normalized probability distribution in variables U and S

$$g_{U,S}(U,S) = \int_{W=-\sqrt{4US}}^{+\sqrt{4US}}\int_{X=-\sqrt{4U-4U^2-4US}}^{+\sqrt{4U-4U^2-4US}} g(U,S,W,X)\,\mathrm{d}W\,\mathrm{d}X = \exp\left[-\frac{U(\epsilon_\beta-\epsilon_\alpha)}{k_BT}\right]\exp\left[-\frac{S(\epsilon_\gamma-\epsilon_\alpha)}{k_BT}\right]\exp\left(-\frac{\epsilon_\alpha}{k_BT}\right) = \exp\left(-\frac{1}{k_BT}\sum_{i=\alpha,\beta,\gamma}|c_i|^2\epsilon_i\right) \quad (23)$$

The probability distribution function $g_{U,S}(U,S)$ is a function of only two independent variables, U and S, that is, $|c_\beta|^2$ and $|c_\gamma|^2$, since the population of the ground state $|c_\alpha|^2$ can be expressed in terms of the other two populations in accordance with conservation of norm. The simple product form of eq 23 appears deceptively simple. We note that the state populations are not independent as a simple product might suggest but are correlated because of the constraints that $0 < |c_i|^2 < 1$ for each state i, and the sum over all states of $|c_i|^2$ is unity.

4. N-Level Quantum Subsystem

It is straightforward to generalize eq 23 to obtain the un-normalized probability distribution for an arbitrary number of quantum levels, N:

$$g_{2,3,..N}(|c_2|^2,|c_3|^2,\ldots|c_N|^2) = \exp\left(-\frac{1}{k_BT}\sum_{i=1}^{N}|c_i|^2\epsilon_i\right) \quad (24)$$

with the constraints $0 < |c_i|^2 < 1$ for each state i, and the sum over all states of $|c_i|^2$ is unity. As for the three-level

case, the probability distribution functions are correlated because of the constraint of conservation of norm.

We now derive the mean energy of an N-state quantum subsystem, using eq 24 for the probability distribution of the populations:

$$\bar{E} = \left\{ \int_0^1 \mathrm{d}|c_2|^2 \int_0^{1-|c_2|^2} \mathrm{d}|c_3|^2 \int_0^{1-|c_2|^2-|c_3|^2} \mathrm{d}|c_4|^2 \ldots \int_0^{1-\sum_{i=2}^{N-1}|c_i|^2} \mathrm{d}|c_N|^2 \left(\sum_{i=2}^{N} \epsilon_i |c_i|^2\right) \exp\left(\frac{-1}{k_B T} \sum_{i=2}^{N} \epsilon_i |c_i|^2\right) \right\} \Bigg/ \left\{ \int_0^1 \mathrm{d}|c_2|^2 \int_0^{1-|c_2|^2} \mathrm{d}|c_3|^2 \int_0^{1-|c_2|^2-|c_3|^2} \mathrm{d}|c_4|^2 \ldots \int_0^{1-\sum_{i=2}^{N-1}|c_i|^2} \mathrm{d}|c_N|^2 \exp\left(\frac{-1}{k_B T} \sum_{i=2}^{N} \epsilon_i |c_i|^2\right) \right\} \quad (25)$$

The ground-state energy ϵ_1 is assumed to be zero for convenience. The integrations limits in eq 25 result from conservation of norm; that is, the constraint that the sum of all probabilities be unity. Equation 25 is the basic result of this paper. The equilibrium mean energy of the quantum subsystem by the Ehrenfest method depends only on the energies of the quantum levels and is independent of the coupling strength. It is also clearly not a Boltzmann distribution, as shown below.

In the special case that the energy levels of the quantum subsystem are nondegenerate and equally spaced, eq 25 can be integrated to obtain a closed-form expression for the mean energy. Assume there are N quantum states with energies 0, ϵ, 2ϵ, ...$(N-1)\epsilon$. The integration of eq 25 gives the simple result

$$\bar{E} = \frac{(N-1)[1 + \epsilon/k_B T - \exp(\epsilon/k_B T)]k_B T}{1 - \exp(\epsilon/k_B T)} \quad (26)$$

Note that the mean energy is proportional to $N-1$, the number of quantum states minus 1. Thus, for an infinite number of states, that is, for the harmonic oscillator, the equilibrium mean energy of the quantum subsystem is infinite at any finite temperature of the classical bath.

5. Simulations

To verify the derived expression for the Ehrenfest mean energy of an N-level system with equal energy spacing, eq 26, we have carried out numerical simulations. The classical subsystem is represented by a linear chain of N_c particles, coupled to each other by anharmonic, nearest-neighbor potentials given by

$$V(\mathbf{q}) = \sum_{k=1}^{N_c} V_M(q_k - q_{k+1}) \quad (27)$$

where

$$V_M(q) = V_0(a^2q^2 - a^3q^3 + 0.58a^4q^4) \quad (28)$$

and q_{N_c+1} is a fixed position. Anharmonic interactions are required to achieve ergodicity, that is, to ensure that the system achieves a true equilibrium independent of the classical and quantum initial conditions. This was verified numerically. The quantum subsystem is an N-level system coupled to atom 1 of the classical chain. The assumption that the quantum subsystem is coupled only to the first atom of the chain is for convenience only; it does not affect any of the conclusions. The quantum energy levels and nonadiabatic couplings are taken to be independent of $\mathbf{q}$. The number of classical atoms in the chain was typically chosen to be 20. A Langevin friction constant γ and white random force $F(t)$ were imposed on atom number N_c of the chain, the one most distant from the quantum subsystem, to ensure that the classical subsystem maintained the correct canonical ensemble equilibrium. $F(t)$ is a Gaussian random variable of width given by[29]

$$\sigma = (2\gamma m k_B T \delta^{-1})^{1/2} \quad (29)$$

where δ is the time step of the integration. The parameters in eqs 28 and 29 were chosen to be the same as those in a previous publication,[21] $V_0 = 175$ kJ/mol, $a = 4.0$ Å^{-1}, $\gamma = 10^{14}$ s^{-1}, and $m = 12$ amu. The classical equations of motion were integrated using a modified Beeman algorithm.[29,30] The quantum equations of motion, describing the time evolution of the complex expansion coefficients of the wave function ψ

$$i\hbar\dot{c}_k = c_k\epsilon_k - i\hbar\sum_{j=1}^{N} \dot{\mathbf{R}}\cdot\mathbf{d}_{kj} \quad (30)$$

were integrated using the fourth-order Runge Kutta algorithm.[31] In our one-dimensional linear chain model, the term $\dot{\mathbf{R}}\cdot\mathbf{d}_{kj}$ is replaced by $p_1 d_{kj}/m$ as discussed earlier. The nonadiabatic couplings d_{kj} between quantum states k and j are given by the expression

$$d_{kj} = \frac{-\epsilon^2\sqrt{\frac{m_H}{2\hbar^2\epsilon}}(\sqrt{j}\,\delta_{k-1,j} + \sqrt{j-1}\,\delta_{k-1,j-2})}{\epsilon_j - \epsilon_k} \qquad j \neq k;\ j = 1, \ldots N;\ k = 1, \ldots N$$

$$d_{kk} = 0 \quad (31)$$

In the above equation, $\delta_{k,j}$ is the Kronecker delta. The energy gap ϵ between adjacent levels was chosen to be 35.9 kJ/mol, and m_H was 1 amu. Ehrenfest simulations were performed for a number of quantum states, $N = 2$, 4, 6, 8, and 10, and for $\epsilon/k_B T$ ranging from 0.54 (high temperature) to 2.16 (low temperature). The equilibrium averages of the Ehrenfest simulations were obtained from an ensemble of 20 trajectories, each typically 50 ps in length, with a time step ≤ 0.005 fs. The initial 20 ps was neglected in the averages to remove any dependence on initial conditions. The same equilibrium populations were obtained whether the quantum system started initially in any pure state or in a linear combination of quantum states, confirming that the system is indeed ergodic and that the averages correspond to true equilibrium averages, within statistical uncertainties. For low temperatures ($\epsilon/k_B T > 2$) and large N values ($N >$

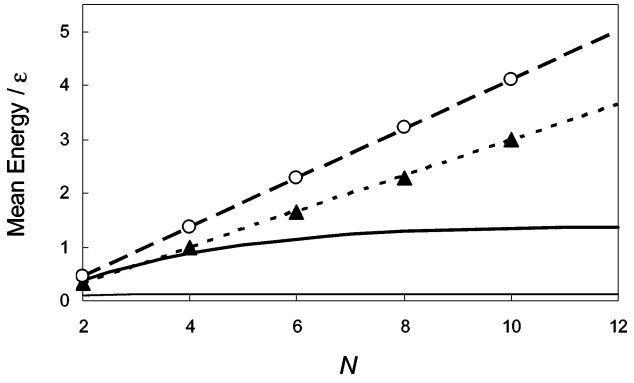


Figure 1. Mean energy of a quantum oscillator with equally spaced energy levels, as a function of the number of quantum states, *N*. The dashed lines are the mean energies obtained by the Ehrenfest method, from eq 26, for $\epsilon/k_BT = 0.54$ (− − −) and $\epsilon/k_BT = 2.16$ (- - -). The circles and triangles are the Ehrenfest mean energies obtained from simulations for $\epsilon/k_BT = 0.54$ and 2.16, respectively, confirming the validity of eq 26. Note that the Ehrenfest mean energy does not converge with increasing *N*. The mean Boltzmann energy as a function of *N* is shown for comparison for $\epsilon/k_BT = 0.54$ (—; heavy line) and $\epsilon/k_BT = 2.15$ (—; light line).

6), the length of each trajectory was in the 200−600 ps range, with the initial 100−400 ps neglected, since a longer time was required to achieve equilibrium.

6. Results

Figure 1 shows the mean energy of an *N*-level quantum system, with equally spaced energy levels, as derived for the Ehrenfest method, eq 26. Also shown are simulation results for *N* ranging from 2 to 10 and for $\epsilon/k_BT = 0.54$ and 2.16. The simulations are in full agreement, within statistical uncertainties, with eq 26. As shown in Figure 1, the Ehrenfest result differs substantially from the desired Boltzmann distribution for an *N*-state quantum system

$$\bar{E}_{\mathrm{BOLTZ}} = \frac{\sum_{i=1}^{N} \epsilon_i \, \mathrm{e}^{-\epsilon_i/k_BT}}{\sum_{i=1}^{N} \mathrm{e}^{-\epsilon_i/k_BT}} \tag{32}$$

The Ehrenfest mean energy does not converge as *N* increases, in contrast to the Boltzmann mean energy, which closely approaches its asymptotic limit by $N = 10$ for the parameters of Figure 1.

$$\bar{E}_{\mathrm{BOLTZ},\infty} = \frac{\epsilon \exp(-\epsilon/k_BT)}{1 - \exp(-\epsilon/k_BT)} \tag{33}$$

The Ehrenfest mean energy also deviates from the classical expression for the mean energy of a simple harmonic oscillator, $\bar{E}_{\mathrm{CL}} = k_BT$. In the limit, as $N \to \infty$, it is clear that, at any nonzero temperature, no matter how low, the mean energy of the Ehrenfest quantum subsystem diverges, $\bar{E} \to \infty$, as shown in Figure 1. This is a nonphysical result.

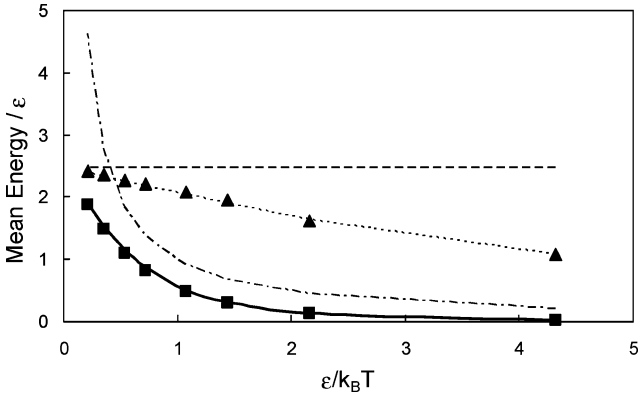


Figure 2. Mean energy of the quantum oscillator with six equally spaced energy levels as a function of inverse temperature. Triangles are the mean energies obtained by the Ehrenfest simulations. The dotted line is the analytical expression, eq 26, which agrees with the Ehrenfest simulations within statistical uncertainties. The horizontal dashed line is the mean energy at infinite temperature, i.e., when all six states are equally populated. The dash-dot shows the classical mean energy of a harmonic oscillator, k_BT. The solid curve shows the energy obtained from a Boltzmann distribution of populations for the six-level system. Squares are the results of fewest-switches surface-hopping simulations, as described elsewhere,[21] showing that surface hopping does achieve the correct Boltzmann equilibrium limit for the quantum subsystem.

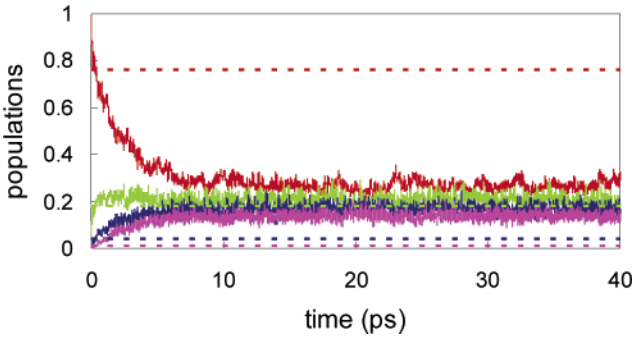


Figure 3. Ensemble averaged populations as a function of time for a quantum oscillator with six equally spaced energy levels, from an Ehrenfest simulation. The simulation was carried out at $\epsilon/k_BT = 1.44$. The quantum subsystem was started in the ground state ($n = 0$). However, the final steady-state populations do not depend on the initial state. Red, green, blue, and magenta correspond to $n = 0$, 1, 2, and 3, respectively. The horizontal lines show the Boltzmann populations at $\epsilon/k_BT = 1.44$. Only the first four levels are shown for clarity.

Figure 2 shows the mean energy of the quantum subsystem, for number of quantum levels $N = 6$, as a function of the unitless energy ϵ/k_BT, that is, as a function of inverse temperature. As shown in Figure 2, the closed-form expression for the Ehrenfest mean energy, eq 26, is further verified by the simulations. The Ehrenfest mean energy is again seen to deviate substantially from both the quantum and classical Boltzmann mean energies. For comparison, the results of fewest-switches surface-hopping simulations, carried out by the procedure of a previous reference,[21] are also shown in

Figure 2. The surface-hopping results agree with the quantum Boltzmann results within statistical uncertainty. This offers additional verification of our previous demonstration that fewest-switches surface hopping satisfies detailed balance rigorously.[21]

Figure 3 shows an Ehrenfest simulation of the time evolution of the quantum state populations for a six-state quantum subsystem of equally spaced levels (only the four lowest-energy levels are shown), with 50 trajectories in the ensemble. It can be seen from Figure 3 that the Ehrenfest state populations, while not equal, are relatively close in magnitude, in contrast to the Boltzmann populations (horizontal dashed lines).

7. Conclusions

We have analyzed the long-time, equilibrium limit of the Ehrenfest MQCD method using a Liouville-like equation for the time evolution of the distribution function of the phase space variables. We find that the Ehrenfest method fails to achieve the correct long-time, equilibrium state; the quantum subsystem does not approach the same temperature as the classical bath that drives it. Rather, the populations of quantum levels are non-Boltzmann, and the mean energy of the quantum subsystem is too high. For the particular case of a quantum oscillator with N equally spaced levels in contact with a bath of an infinite number of classical particles, we have derived a simple closed-form analytical expression for the equilibrium mean energy of the quantum subsystem. We have verified this expression via simulations. The equilibrium Ehrenfest mean energy can be significantly higher than that given by the Boltzmann distribution of populations and nonphysically diverges with increasing N. The quantum subsystem does not approach infinite temperature in this limit; the level populations decrease with increasing energy, but do so sufficiently slowly so that the mean energy diverges.

The failure of the Ehrenfest method to achieve the correct thermal equilibrium is a result of the fact that the squares of the quantum amplitudes, $|c_i|^2$, which are the phase space variables for the quantum subsystem, are continuous, classical-like variables. Thus, the expectation value of an observable requires integration over $\mathrm{d}|c_i|^2$, rather than a discrete sum over i, that is, over the diagonal elements of the density matrix. As the number of states N increases, the number of integration variables $\mathrm{d}|c_i|^2$ in the Ehrenfest theory increases proportionally, whereas the correct quantum result remains a sum over a single variable i. The consequence of this is that the effective volume of phase space corresponding to a particular energy E increases nonphysically with increasing E, giving rise to a higher average energy than the correct quantum Boltzmann result.

Failure of Ehrenfest MQCD to achieve the correct equilibrium state may seriously limit its applicability. Certainly, this prohibits its use to compute equilibrium properties. In addition, applications of the Ehrenfest method to study the rate of approach to equilibrium, such as energy relaxation, solvent reorganization, or nonradiative decay, must be carried out with caution. More generally, detailed balance relates the equilibrium populations of two states to the ratio of forward and backward rates. If detailed balance is not satisfied, then at least one of the rates must be in error, possibly affecting even short time dynamics. As demonstrated previously,[21] the alternative, widely used MQCD method, fewest-switches surface hopping, does not suffer from this deficiency; at long times, the quantum and classical subsystems rigorously approach the same temperature.

Acknowledgment. This work was supported by the National Science Foundation, Grant CHE0314208.

References

(1) Nikitin, E. E. *Theory of Elementary Atomic and Molecular Processes in Gases*; Clarendon Press: Oxford, U. K., 1974.

(2) Jasper, A. W.; Zhu, C.; Nangia, S.; Truhlar, D. G. *Faraday Discuss.* **2004**, *127*, 1.

(3) Hammes-Schiffer, S.; Tully, J. C. *J. Chem. Phys.* **1994**, *101*, 4657.

(4) Staib, A.; Borgis, D.; Hynes, J. T. *J. Chem. Phys.* **1995**, *102*, 2487.

(5) Consta, S.; Kapral, R. *J. Chem. Phys.* **1996**, *104*, 4581.

(6) Billeter, S. R.; Webb, S. P.; Iordanov, T.; Agarwal, P. K.; Hammes-Schiffer, S. *J. Chem. Phys.* **2001**, *114*, 6925.

(7) Tully, J. C. In *Modern Methods for Multidimensional Dynamics Computations in Chemistry*; Thompson, D. L., Ed.; World Scientific: Singapore, 1998; p 34.

(8) Pechukas, P. *Phys. Rev.* **1969**, *181*, 174.

(9) Herman, M. F. *Annu. Rev. Phys. Chem.* **1994**, *45*, 83.

(10) Prezhdo, O. V.; Brooksby, C. *Phys. Rev. Lett.* **2001**, *86*, 3215.

(11) Burant, J. C.; Tully, J. C. *J. Chem. Phys.* **2001**, *112*, 6097.

(12) Kernan, D. M.; Ciccotti, G.; Kapral, R. *J. Chem. Phys.* **2002**, *116*, 2346.

(13) Tully, J. C. *J. Chem. Phys.* **1990**, *93*, 1061.

(14) Tully, J. C.; Preston, R. K. *J. Chem. Phys.* **1971**, *55*, 562.

(15) Tully, J. C. In *Modern Theoretical Chemistry: The Dynamics of Molecular Collisions*; Miller, W. H., Ed.; Plenum Press: New York, 1976; p 217.

(16) McLachlan, A. D. *Mol. Phys.* **1964**, *8*, 39.

(17) Micha, D. A. *J. Chem. Phys.* **1983**, *78*, 7138.

(18) Kirson, Z.; Gerber, R. B.; Nitzan, A.; Ratner, M. A. *Surf. Sci.* **1984**, *137*, 527.

(19) Sawada, S. I.; Nitzan, A.; Metiu, H. *Phys. Rev. B* **1985**, *32*, 851.

(20) Billing, G. D. *The Quantum Classical Theory*; Oxford University Press: Oxford, U. K., 2003.

(21) Parandekar, P. V.; Tully, J. C. *J. Chem. Phys.* **2005**, *122*, 094102.

(22) Käb, G. *J. Phys. Chem. A.* **2004**, *108*, 8866−8877.

(23) Käb, G. *Phys. Rev. E* **2002**, *66*, 046117.

(24) Mauri, F.; Car, R.; Tosatti, E. *Europhys. Lett.* **1993**, *24*, 431−436.

(25) Meyer, H. D.; Miller, W. H. *J. Chem. Phys.* **1979**, *70*, 3214.

(26) Tully, J. C. *Faraday Discuss.* **1998**, *110*, 407.

(27) Larsen, R. E.; Parandekar, P. V.; Tully, J. C. Unpublished work.

(28) McQuarrie, D. A. In *Statistical Thermodynamics*; University Science Books: Mill Valley, California, 1973; p 117.

(29) Tully, J. C.; Gilmer, G. H.; Shugard, M. *J. Chem. Phys.* **1979**, *71*, 1630.

(30) Beeman, D. *J. Comput. Phys.* **1976**, *20*, 130.

(31) Press, W. H.; Teukolsky, S. A.; Vellerling, W. T.; Flannery, B. P. *Numerical Recipes in C*, 2nd ed.; Cambridge University Press: New York, 1992.

CT050213K

JCTC Journal of Chemical Theory and Computation

Path Integral Simulations of Proton Transfer Reactions in Aqueous Solution Using Combined QM/MM Potentials

Dan Thomas Major, Mireia Garcia-Viloca,† and Jiali Gao*

Department of Chemistry and Supercomputing Insititute, Digital Technology Center, University of Minnesota, Minneapolis, Minnesota 55455

Received October 18, 2005

Abstract: A bisection sampling method was implemented in path integral simulations of chemical reactions in solution in the framework of the quantized classical path approach. In the present study, we employ a combined quantum mechanical and molecular mechanical (QM/MM) potential to describe the potential energy surface and the path integral method to incorporate nuclear quantum effects. We examine the convergence of the bisection method for two proton-transfer reactions in aqueous solution at room temperature. The first reaction involves the symmetrical proton transfer between an ammonium ion and an ammonia molecule. The second reaction is the ionization of nitroethane by an acetate ion. To account for nuclear quantum mechanical corrections, it is sufficient to quantize the transferring light atom in the ammonium ion-ammonia reaction, while it is necessary to also quantize the donor and acceptor atoms in the nitroethane-acetate ion reaction. Kinetic isotope effects have been computed for isotopic substitution of the transferring proton by a deuteron in the nitroethane-acetate reaction. In all computations, it is important to employ a sufficient number of polymer beads along with a large number of configurations to achieve convergence in these simulations.

Introduction

The incorporation of nuclear quantum mechanical effects into simulations of chemical reactions in solution and in enzymes is a challenging task because it is necessary to average over protein conformations and solvent configurations.[1−3] These effects, including zero-point energy and tunneling, are particularly significant for proton and hydride transfer reactions, which are ubiquitous in chemical and enzymatic processes.[1] A widely used approach to probe quantum mechanical tunneling is through measurements of primary and secondary kinetic isotope effects.[4] For example, in two of the most extensively studied enzyme reactions,[5−8] the hydride transfer reaction by liver alcohol dehydrogenase[5,6,9,10] and the proton-transfer reaction by methylamine dehydrogenase[7,8,11] have been shown to have significant tunneling contributions. Both experimental and computational studies suggest that tunneling makes little contributions to *catalysis*,[2] which is related to the rate enhancement by an enzyme relative to the uncatalyzed process in water. It is, nevertheless, essential to include quantum mechanical effects to determine kinetic isotope effects and to estimate the reaction rates quantitatively.[1,3,12] Furthermore, a number of theoretical studies have shown that the inclusion of zero point energy can reduce free energy barriers by 2−3 kcal/mol for enzyme reactions.[10,13,14]

Several simulation methods have been used to determine kinetic isotope effects in enzymatic reactions, including the ensemble-averaged variational transition state theory with multidimensional tunneling (EA-VTST),[9] discretized path integral simulations,[15−17] and a multiconfiguration wave function method.[10] These methods have been applied to several enzymatic reactions with good accord between the calculated and experimental kinetic isotope effects.[1] The EA-VTST approach also has the advantage of separating contributions from bound vibrations and tunneling, providing further insights into the reaction mechanism. In this paper,

* Corresponding author e-mail: gao@chem.umn.edu.

† Present address: Institute of Biotechnology and Biomedicine, Universitat Autònoma de Barcelona, Bellaterra 08193, Catalunya, Spain.

10.1021/ct050257t CCC: $33.50
Published on Web 01/26/2006

we describe the implementation of a bisection sampling algorithm in centroid path integral simulations and examine the convergence properties in these calculations for two proton-transfer reactions in solution;[18−20] the first system is a model reaction of $[H_3N-H-NH_3]^+$ and the other is the proton abstraction of nitroethane by acetate ion, modeled for the enzymatic process in nitroalkane oxidase. We implement an approach similar to that described by Hwang et al.[15,16] and by Sprik et al.,[21] in which the quantum mechanical effects are incorporated into the rate calculation through a transmission coefficient by correcting the classical potential of mean force (PMF) obtained from Monte Carlo or molecular dynamics simulations.[2,3,22] Thus

$$k^{qm} = \gamma k^{TST} \tag{1}$$

where k^{qm} is the quantum mechanical rate constant and k^{TST} is the classical transition state theory (TST) rate constant. In general, the transmission coefficient in eq 1 is a product of the deviation from equilibrium behavior, the classical dynamic recrossing factor, Γ, and the quantum mechanical correction, κ.[2,3] Here, we focus on the quantum mechanical contributions, which are defined as follows[15,16]

$$\kappa = e^{-\beta(G^{\ddagger}_{qm} - G^{\ddagger}_{TST})} \tag{2}$$

where $\beta = 1/k_B T$, k_B is Boltzmann's constant, T is the temperature, and $G^{\ddagger}_{qm}$ and $G^{\ddagger}_{TST}$ are the quantum and classical free energy of activation, respectively.

Although the quantum mechanical free energy of activation, $G^{\ddagger}_{qm}$, can be obtained directly by using centroid path integral molecular dynamics simulations,[23−27] Hwang et al. noted that it is more convenient to evaluate the free energy difference, $G^{\ddagger}_{qm} - G^{\ddagger}_{TST}$, in eq 2.[15,16] Thus, rather than carrying out a full centroid path integral simulation directly, one performs classical molecular dynamics simulations to obtain the potential of mean force along a reaction coordinate, and then a quantum correction is made along the classical reaction path.[15,28] This provides a quantum correction to the classical results. A similar idea was originally described by Sprik et al.,[21] who proposed a procedure to obtain quantum mechanical averages through free-particle path integral sampling over classical configurations from molecular dynamics or Monte Carlo simulations.

$$\langle A \rangle = \langle \langle A \rangle_K^{FP} \rangle_{CM} \tag{3}$$

In eq 3, the inner average $\langle A \rangle_K^{FP}$ represents the quantum average of property A by path integral sampling of free-particles over a fixed configuration K, in which the center of mass (centroid) positions are constrained to those of the classical particle positions. The outer average is over "classical" configurations. This double averaging strategy is the essence of the quantized classical path (QCP) method exploited by Hwang et al.,[15] which employs the trajectory obtained from classical molecular dynamics simulations to obtain the QM correction by performing free-particle path integral averaging. The QCP method can be used to treat nuclear QM effects in macromolecular systems, and it has been applied to several enzymatic reactions.[15,17,29,30]

A central issue in path integral simulations is convergence. It appears that a direct sampling procedure was used in previous QCP applications to enzymatic reactions. Åqvist and co-workers used the QCP approach to calculate the kinetic isotope effect in the proton-transfer reaction catalyzed by glyoxalase I.[17] In this study, 20 particle-beads were used to describe the quantized paths, and for each classical configuration, 1, 5, and 10 Monte Carlo Metropolis steps were used, respectively, to obtain the quantum corrections. In another calculation, a total of 20 000 free-particle configurations were used for 18 beads along the entire reaction coordinate for an enzymatic reaction.[30] Other studies indicate that more extensive path integral sampling might be needed even for a dilute hard-sphere system.[21] In the present study, we use a bisection free-particle sampling method,[19] coupled with the QCP approach to enhance convergence.[15] The convergence properties of the method are scrutinized with a view to arrive at a practical scheme for condensed phase simulations. Here, we present results for the two model proton-transfer reactions in aqueous solution mentioned above. The first reaction is the symmetric proton transfer between an ammonium ion and an ammonia molecule (reaction I). The second reaction is between nitroethane and an acetate ion (reaction II)

$$NH_4^+ + NH_3 \rightarrow NH_3 + NH_4^+$$

$$CH_3CO_2^- + CH_3CH_2NO_2 \rightarrow CH_3CO_2H + CH_3\bar{C}HNO_2$$

The convergence of the QM correction to the classical PMF is analyzed with respect to the sampling of the path-integrals, the solution phase classical configurations, and the number of ring polymer beads. The conclusions from the current work will be useful in future simulations of solution phase and enzymatic reactions.

Theoretical Background

The centroid quantum mechanical partition function for a ring of P quasi-particles or beads in the discrete path integral form is given as follows[31]

$$Q_P^{qm} = \int d\bar{\mathbf{x}} \left(\frac{1}{4\pi\Lambda^2}\right)^{P/2} \int dx_1 \cdots \int dx_P e^{-\beta V^{qm}} \tag{4}$$

where $\beta = 1/k_B T$, P is the number of quasi-particles of the discrete path, and the average or centroid position, $\bar{\mathbf{x}}$, of the quasi-particle positions, $\{x_i; i = 1, \cdots P\}$, is defined as

$$\bar{\mathbf{x}} = \frac{1}{P}\sum_{i=1}^{P} x_i \tag{5}$$

In eq 4, the effective potential $V^{qm}(x_1, \cdots, x_p)$ is given by

$$V^{qm}(x_1, \cdots, x_p) = \frac{1}{4\beta\Lambda^2}\sum_{k}^{P}(x_k - x_{k+1})^2 + \frac{1}{P}\sum_{k}^{P} U(x_k) \tag{6}$$

and Λ is the thermal de Broglie wavelength

$$\Lambda = \left(\frac{\beta\hbar^2}{2mP}\right)^{1/2} \tag{7}$$

where m is the mass of the particle. Thus, each quasiparticle is connected by a harmonic spring with its two neighbors and is subjected only to a fraction, $1/P$, of the full classical potential, $U(x_i)$. The discrete paths are circular with $x_{P+1} = x_1$. The exact QM partition function is obtained in the limit

$$Q^{\mathrm{qm}} = \lim_{P\to\infty} Q_P^{\mathrm{qm}} \tag{8}$$

In the quantized classical path (QCP) approach,[15] Warshel and co-workers showed that the QM correction to classical free energy along a reaction path can be determined by a double average of classical and free-particle path integral simulations, making use of the assumption that the centroid positions coincide with the classical coordinates.[24,32] Thus

$$G^{\ddagger}_{\mathrm{qm}}(\bar{x}) - G^{\ddagger}_{\mathrm{TST}}(\bar{x}) = -\frac{1}{\beta}\ln\frac{Q_P^{\mathrm{qm}}}{Q_P^{\mathrm{cm}}} = -\frac{1}{\beta}\ln \ll e^{-\beta/P\sum_k^P \Delta U_k} >_{\mathrm{FP},\bar{x}} >_{U(\bar{x})} \tag{9}$$

where Q_P^{cm} is the reference classical (CM) partition function, and $\Delta U_k = U(x_k) - U(\bar{x})$. Here, the outer average $< \cdots >_{U(\bar{x})}$ is obtained according to the distribution generated by propagating classical molecular dynamics or Monte Carlo simulations using the potential $U(\bar{x})$. The inner average $< \cdots >_{\mathrm{FP},\bar{x}}$ is over the free particle distribution, in the absence of any external potential[15]

$$< e^{-\beta/P\sum_k^P \Delta U(x_k)} >_{\mathrm{FP},\bar{x}} = \frac{\int \delta(\bar{x}) e^{-\beta/P\sum_k^P \Delta U(x_k)} e^{-1/4\beta\Lambda^2\sum_k^P (x_k - x_{k+1})^2} dx_1 \cdots dx_P}{\int \delta(\bar{x}) e^{-1/4\beta\Lambda^2\sum_k^P (x_k - x_{k+1})^2} dx_1 \cdots dx_P} \tag{10}$$

where the integration of beads is constrained at the centroid position $\bar{\mathbf{x}}$. The advantage of this formulation is that one can sample the free particle (FP) distribution (i.e. the quasiparticle polymer rings) separately at each CM configuration (i.e. centroid position) and then average over all CM configurations obtained from molecular dynamics simulations.

Computational Details

A. Convergence. Although eqs 9 and 10 provide a very appealing framework for obtaining quantum averages by carrying out classical simulations on a classical potential $U(\bar{x})$, a main practical problem is that most configurations obtained from the free-particle sampling procedure (inner average of eq 9) have very small contributions to the total average in the external potential $U(\bar{x})$. Only a tiny fraction of the free particle configurations have sufficiently large probabilities in the actual physical environment. Thus, unless an efficient free particle sampling procedure is used, it would be very difficult to achieve convergence using eq 10.

In principle, it is possible to find a subset of free particle configurations that carry the greatest weight to the path integral average in QCP calculations. However, for a given potential energy surface, these free particle configurations are generally not known, unless the potential energy surface and Hessian for each classical configuration have been enumerated.[27] Therefore, it is necessary to comprehensively sample the free particle distribution. Moreover, to ensure convergence from PI calculations, it is necessary to increase the number of beads until the desired property is converged. We have often encountered diverging averages in the quantized classical path approach as the number of particle beads is increased if a direct sampling procedure is used for the free particle distributions. This is because the spring connecting the polymer beads becomes increasingly stiff as the number of beads increases, which makes it more difficult to sample. In fact, the equilibration time of the slowest mode in the ring polymer scales as $(P/\pi)^2$ where P is the number of beads.[19,20] It has been noted that standard Metropolis Monte Carlo and molecular dynamics simulations are not the optimal choice as a free particle sampling algorithm.[33]

In the present study, we employed the bisection scheme introduced by Ceperley and co-workers,[19,20] which turns out to be the most effective method in our application. Here, the free-particle distribution can be sampled exactly because it is a Gaussian of known mean and width

$$\rho(x_i, x_m; \beta/P) = \left(\frac{1}{4\pi\sigma}\right)^{1/2} e^{(x_i - x_m)^2/2\sigma} \tag{11}$$

where the variant of the Gaussian is the square of the de Broglie wavelength $\sigma = \Lambda^2$, and

$$x_m = \frac{(x_{i-1} + x_{i+1})}{2} \tag{12}$$

We have implemented the bisecting method proposed by Ceperley in QCP calculations,[19] which is a combination of multilevel Monte Carlo[20] and the Lévy construction for sampling a free particle path.[34] The bisection method takes advantage of the fact that the density matrix at a given temperature may be written as the integral of two density matrices at a higher temperature.[19] Thus one can accurately sample the free-particle distribution at a higher temperature and more effectively explore the configurational space. The present bisection quantized classical path sampling is called BQCP.[18]

Specifically, for each free particle move, we select a random sequence of N-1 consecutive beads in the polymer ring, where $N = 2^l$ and l is called the level of bisection.[19] The ends of the bead sequence are fixed at r_i and r_{i+N}. In principle, the two endpoints could be the same bead, when the entire polymer ring is sampled at each step, which will generate entirely uncorrelated configurations. At the coarsest level of bisection, $k = l$, the position of the bead in the midpoint of the sequence is first sampled, which is placed at the geometrical center of the two end points and randomly displaced according to the Gaussian distribution of width $2^{l-1}\sigma$, $r_{i+N/2} = (r_i + r_{i+N})/2 + \xi$, where ξ is the random displacement vector. Having sampled the $r_{i+N/2}$ point, we bisect the two new intervals, $(r_i, r_{i+N/2})$ and $(r_{i+N/2}, r_{i+N})$ at the next bisection level $k = l - 1$ with the distribution width $2^{l-2}\sigma$, to sample points at $r_{i+N/4}$ and $r_{i+3N/4}$. This bisecting procedure continues recursively until level $k = 1$, where all $N-1$ beads have been sampled. As in the single bead sampling, the acceptance ratio for such a "Monte Carlo"

Table 1. Reaction and Transition State[a] Energies for Reactions I and II

	AM1	AM1-SRP	ab initio
reaction I	0.0 (3.3)	ND	0.0 (2.6)[b]
reaction II	−9.6 (3.1)	7.5 (11.7)	9.8[c](12.5)[d]

[a] Transition state energies in parentheses. [b] Reference 51. [c] Reference 45. [d] MP2/6-31+G(d)// MP2/6-31+G(d).

move will be 100% since the new positions are drawn from the accurate free-particle distribution.[19] After each move, the polymer ring is recentered at the classical position to enforce the centroid constraint. In our implementation, the Box-Muller transformation is used to generate the random displacements with a Gaussian distribution.[35]

The BQCP method has been implemented in the CHARMM simulation package[36] in version c32a2 in a serial and a parallel version. The quantum correction to the classical PMF curves were fitted to an inverse Eckart potential using the Levenberg−Marquardt nonlinear optimization method.[35] The QM correction along the reaction coordinate is assumed to be a smooth, continuous function.[18]

B. Potential Energy Function. To describe nuclear quantum effects in aqueous phase reactions, it is essential to employ a potential energy function that can describe the bond breaking and formation process. We adopted a strategy that combines a quantum mechanical model with a molecular mechanical force field, or combined QM/MM potential, in molecular dynamics simulations. Thus, the solute molecules are treated by quantum mechanics, and the solvent is represented by the three-point charge TIP3P model for water.[37] The total energy of the system is

$$E_T = E_{QM} + E_{QM/MM} + E_{MM} \tag{13}$$

where E_{QM} is the solute energy, E_{MM} is the solvent energy, while $E_{QM/MM}$ is the energy term for interactions between QM and MM atoms. Combined QM/MM methods have been used to study a variety of chemical reactions in the gas-phase, and in condensed phases, including enzymes, and have been described in a number of review articles.[38−42]

In the proton transfer between two ammonia molecules the standard AM1 semiempirical Hamiltonian was employed, as it has been shown to give reliable results for this system (Table 1). However, for the nitroethane-acetate reaction, neither the AM1 nor the PM3 method yielded satisfactory energetic results in comparison with ab initio data at the G3 level. Thus, we developed a set of reaction specific parameters (AM1-SRP), using AM1 as a starting point for a full nonlinear optimization of the parameter space.[43] The parametrization was performed in a stepwise manner, at each step allowing additional parameters to be optimized simultaneously. The process commenced with the U_{ss}, U_{pp}, β_s, and β_p parameters allowing changes up to 15% from the original AM1 values. At the second step the ζ_s, ζ_p, and α parameters were also allowed to change up to 10%, while the subsequent step allowed the Gaussian terms L and M to change up to 5%, followed by the K Gaussian terms, which were also allowed to change by the same amount. At the final level of optimization, the 2 electron terms G_{ss}, G_{sp}, G_{pp}, G_{p2}, and H_{sp} were allowed to change up to 2.5%. Thereafter, the parameters were further optimized to fine-tune the fit to the target data. Only gas-phase molecular descriptors for the reactant and product states were employed as target data in the parametrization process. The descriptors used were molecular heats of formation available for three of the four reactant/product species (acetic acid, acetate, and nitroethane). The remaining heat of formation (nitroethyl anion) was obtained from the three other heats of formation and the computed G3 reaction energy. Additionally, MP2/6-31+G(d) bond distances and angles as well as selected frequencies were used as target data. To allow for an accurate description of the charge distribution in the molecules, Mülliken charges and dipole moments (for the neutral species) computed at the MP4/6-31+G(d) level were used as target data; however, the charge restraint was used primarily to ensure balanced charge polarization rather than strictly fitting these charges. Such a careful parametrization protocol yields an inexpensive, yet highly accurate Hamiltonian that is suitable for QM/MM simulations. The final AM1-SRP yielded a reaction energy of 8.0 kcal/mol in good agreement with the G3 value of 9.8 kcal/mol (Table 1),[44] and these parameters have been given in ref 45.

C. Simulation Details. The solutes were embedded in cubic boxes of water molecules. In the case of reaction I the size of the box was 25 Å^3 giving a total of 502 water molecules. In reaction II, the system dimensions were ∼30 Å^3, resulting in 898 water molecules. Internal water bond distances were constrained to the experimental value using the SHAKE algorithm in all simulations.

In reaction I a spherical group based cutoff of 14 Å was used for both van der Waals and electrostatic interactions. However, for reaction II we employed the particle-mesh Ewald summation method for QM/MM simulations[46] to obtain high-accuracy results enabling direct comparison with experimental data. In these simulations, the van der Waals cutoff was group-based and set to 9.5 Å. All simulations employed molecular dynamics propagated using the leapfrog Verlet algorithm with a 1 fs time step.[47] Periodic boundary conditions were used together with the canonical ensemble (NVT) for reaction I and the isobaric−isothermal ensemble (NPT) for reaction II, both at 25 °C. For each simulation (or window, see below), 100 ps of equilibration was first carried out, which was followed by averaging for 100 ps.

The potential of mean force (PMF) profiles were obtained using the umbrella sampling technique.[48,49] According to this approach the reaction is divided into a series of windows, and in each window a biasing potential (umbrella potential) is applied to allow the reaction to climb over the barrier within the time frame of molecular dynamics simulations. The effect of the biasing potential is subsequently removed when the separate windows are combined to produce the PMF profile. This was done using the weighted histogram analysis method.[50] In the current simulations, between 7 and 15 windows were used to span the reaction coordinates for reactions I and II. The reaction coordinates were defined as the difference between the breaking and forming bonds.

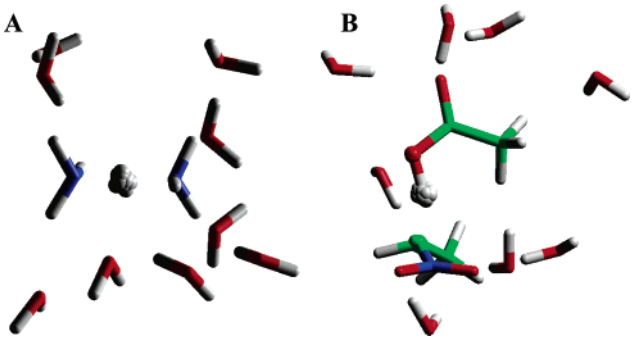


Figure 1. Snapshots of instantaneous structures from combined QM/MM molecular dynamics and quantized classical path simulations for (A) the proton transfer in $NH_4^+ + NH_3$ and (B) the deprotonation of nitroethane by an acetate ion in aqueous solution.

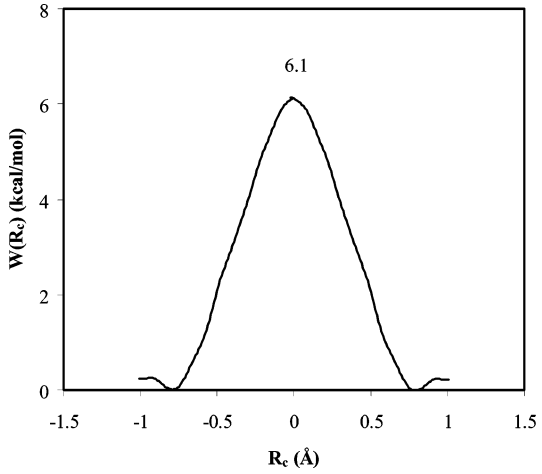


Figure 2. Classical potential of mean force for the NH_4^+-NH_3 proton-transfer reaction in aqueous solution.

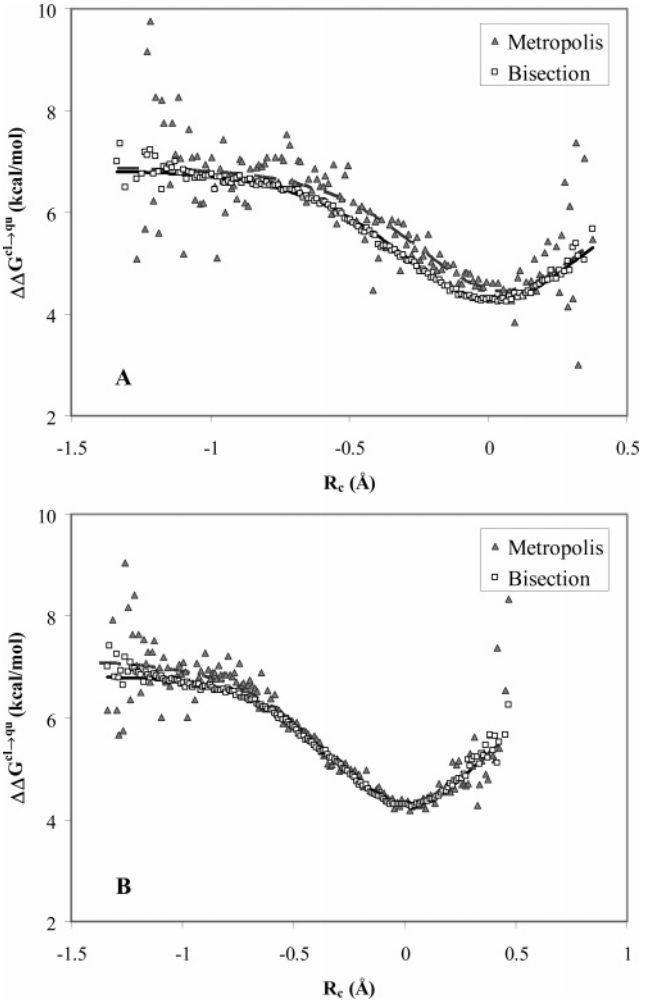


Figure 3. Comparison of the nuclear QM correction for the NH_4^+-NH_3 proton-transfer reaction in aqueous solution using the QCP method with the Metropolis and Bisection sampling schemes.

Results and Discussions

A. Proton Transfer between NH_4^+ and NH_3. The computed classical free energy barrier for reaction I (Figure 1A) is in reasonable agreement with previous results of Garcia-Viloca et al.,[13] although the currently computed barrier of 6.13 kcal/mol (Figure 2) is lower than in that work (11 kcal/mol). The reason for the difference is the use of a fixed nitrogen distance of 3 Å in the previous study,[13] whereas this geometrical parameter was not constrained in the present simulations. This view is supported by two-dimensional PMF simulations performed for this reaction, which showed that the optimal heavy atom distance for proton transfer is ca. 2.6 Å.[51]

We estimated the quantum correction to the classical reaction profile for the proton transfer reaction using the classical trajectories that were saved every 100 integration steps. Initially it is of interest to compare the performance of the QCP method between the standard Metropolis sampling and the bisection sampling of free particles. The use of standard Metropolis sampling with QCP for enzymatic systems has been shown to yield excellent results even with a small number of configurations.[17] However, we were unable to obtain converged results with standard Metropolis sampling of the free particles.[19] Nevertheless, we found that the QCP method coupled with the bisection algorithm converges quickly and obtained reasonable results for model systems.[18] Thus, a systematic comparison between the different sampling schemes for a chemical reaction in the condensed phase is presented. The result of such a comparison of the two sampling methods for reaction I is shown in Figure 3. This test case used 16 beads for each of the three atoms directly involved in the proton transfer (N−H- - -N), similar to the number of beads used in related PI studies.[17] A total of 7000 classical configurations (corresponding to 1000 sets of coordinates from each of the 7 windows used in the classical umbrella sampling simulations, spanning 700 ps) saved at 0.1 ps intervals were used, and each classical point was subjected to 100 MC steps of free particle path integral sampling. This scheme is denoted by (7K/100). It is evident from the results that the use of an accurate free particle distribution greatly improves the convergence of the QM correction to the classical potential energy surface. Using the bisection sampling scheme, a QM correction of 2.53 kcal/mol is obtained. Note that we use two decimal digits in the discussion purely for the purpose of comparing convergence as the overall statistical errors in the computed potential of mean force are about 0.5 kcal/mol based on experience from umbrella sampling simulations starting from different initial

configurations and conditions. The reaction coordinate position of the maximum QM barrier correction is located at 0.01 Å, close to the expected value of zero for a symmetric reaction and is within the statistical accuracy of the data averaging. The convergence of the bisection results was tested by increasing the number of classical steps to 35 000 (35K/100). The results were found to be within ±0.005 kcal/mol, and the position of the maximum QM barrier correction is located at 0.01 Å (Figure 3B). Using the direct Metropolis procedure with a 7K/100 sampling scheme, a QM correction of 2.41 kcal/mol is obtained, with a maximum correction located at 0.05 Å (Figure 3A). Considering that the polymer chains were equilibrated for 1 000 000 steps prior to data collection, we attribute the small difference to incomplete convergence. This conclusion is corroborated by inspecting the χ^2 values obtained from the nonlinear fitting of an inverse Eckart potential to the QM correction curves. Using the bisection sampling scheme, a χ^2 value of 26 is obtained, while with Metropolis sampling χ^2 is 1198. After increasing the number of classical steps to 35000 (35K/100), the QM correction was found to be 2.77 kcal/mol, and the position of the maximum QM barrier correction is located at 0.02 Å. Thus, the results have not yet converged, and additional Metropolis sampling of the PI would be necessary to obtain converged results.

To further probe the convergence behavior of the BQCP method with respect to the number of quantized atoms and the extent of sampling, additional tests were performed for reaction I. These tests used 32 beads, as this has been found to be a reasonable compromise between computational cost and accuracy.[18] The results for three levels of quantization are shown in Figure 4, where 7000 classical configurations were used in conjunction with 100 path integral steps per classical configuration (7K/100). At the lowest level of quantization, only the transferring hydrogen is treated as a ring polymer, while the remaining atoms are classical entities. At the second level, the three transferring atoms are treated quantum mechanically. At the final level, all solute atoms are treated by PI simulations. The quantum mechanical correction obtained when quantizing the transferring proton only is 2.72 kcal/mol, and when quantizing the donor and acceptor heteroatoms as well, the correction term is very similar at 2.67 kcal/mol. Quantizing all solute atoms introduces considerable challenge, due to the large number of hydrogens, which have long de Broglie wavelength and are therefore more difficult to sample. For this particular system, convergence is difficult to achieve when all solute atoms are quantized. With the 7K/100 scheme, a QM correction of 2.81 kcal/mol is obtained. With additional sampling of the classical configuration, using a 14K/100 sampling scheme, the QM correction is 2.59 kcal/mol. It seems that the free particle distribution has not yet been sufficiently sampled, and it is necessary to perform more extensive path integral sampling or to add additional classical configurations.

For reaction I, quantization of only the transferring atom is sufficient to yield reliable results. It is clear from the results that as the number of quantized particles increases, so does the complexity of the ring polymer configurational space.

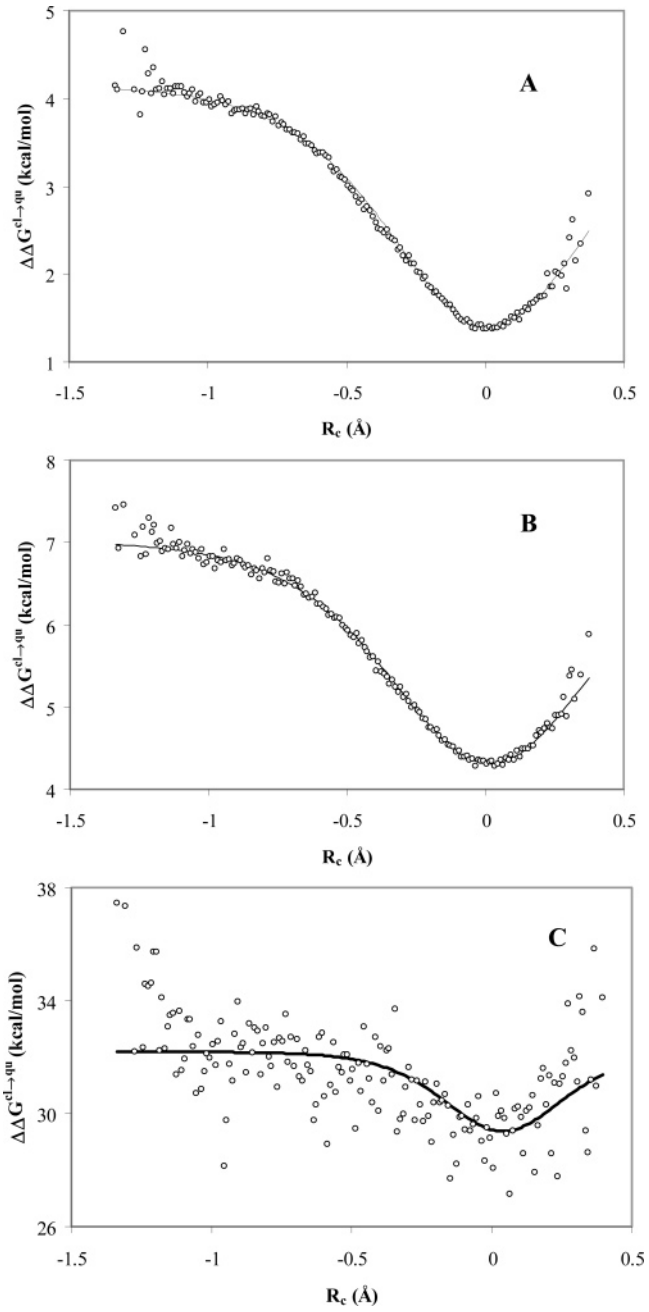


Figure 4. Comparison of the BQCP correction for the NH_4^+-NH_3 proton-transfer reaction in aqueous solution at different levels of quantization: (A) the transferring proton only, (B) the transferring proton plus the donor and acceptor nitrogen atoms (N−H- - -N), and (C) the entire solute (NH_4^+-NH_3).

Thus, quantizing a single atom yields rapid convergence while extending the nuclear QM region slows down the convergence. In a thorough investigation, Tuckerman and Marx showed that the quantum nature of the heavy atoms can substantially enhance proton tunneling by as much as 31% compared to a classical frame in the case of malonaldehyde.[52] Similar findings have also been noted by Hinsen and Roux on acetylacetone.[53] To complement previous convergence tests for simple model systems,[18] we also tested several different sampling schemes to arrive at one that yields the optimal compromise between accuracy and computational cost. In all of the following test cases, the three central atoms

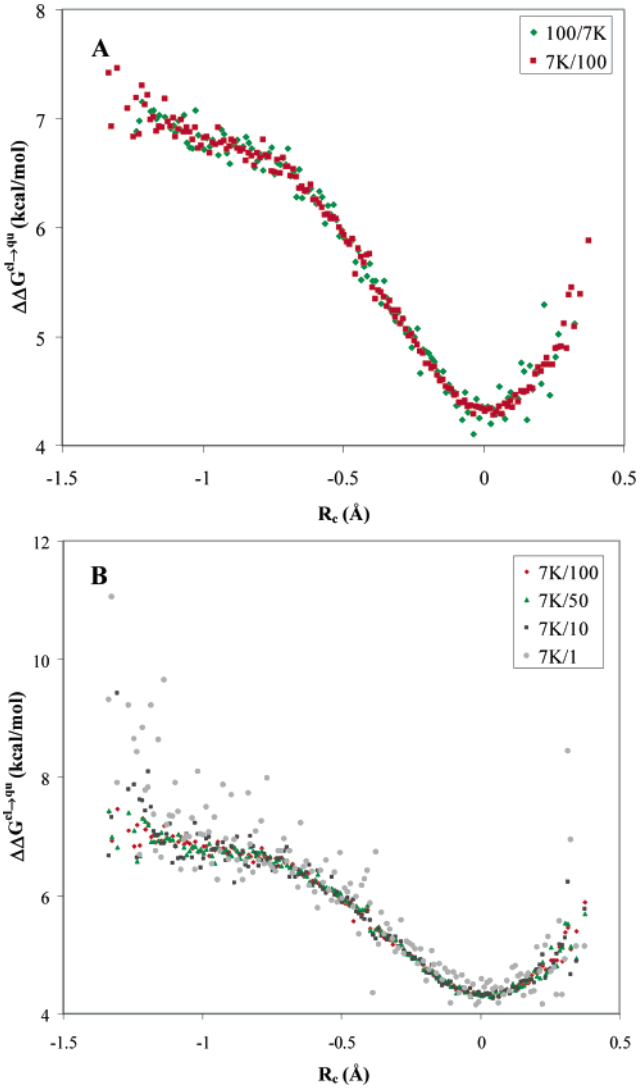


Figure 5. Comparison of different BQCP sampling schemes for the NH_4^+-NH_3 proton-transfer reaction in aqueous solution.

directly involved in the proton transfer were quantized using 32 beads unless otherwise stated.

Previous studies have shown that for a simple Morse potential, approximately 1000 classical configurations coupled with 100 PI steps per classical (centroid) configuration, yields good convergence.[18] Thus, we used this conclusion as a starting point in the present test. In Figure 5A, two sampling schemes, requiring comparable computational cost, are compared. In the first scheme, 7000 classical configurations are coupled with 100 path-integral sampling steps. The second scheme uses 700 classical configurations and 1000 PI steps. Fitting of an inverse Eckart potential to the curves reveals identical QM corrections of 2.67 kcal/mol. However, from the distribution of the points it is clear that a scheme that samples the classical configuration space more extensively is preferable to additional sampling of the polymer rings at each centroid position. This is due to the greater variance in the free particle path-integral estimator (eq 10) than in the external average over the free-particle paths (eq 9).

In Figure 5B, additional such sampling schemes are presented, using 7000 classical configurations, and 100, 50, 10, and 1 PI steps. Although the results indicate that it would be preferable to use 100 PI steps per classical configuration, using 50 or even 10 yields reasonable results. The QM corrections obtained by sampling 100, 50, 10, and 1 quasi-particle configuration(s) at each classical step are 2.67, 2.64, 2.63, and 2.71 kcal/mol, respectively (Figure 5B). The position of the maximum QM correction is 0.01 Å for the former three, whereas for the latter the location has shifted to 0.02 Å.

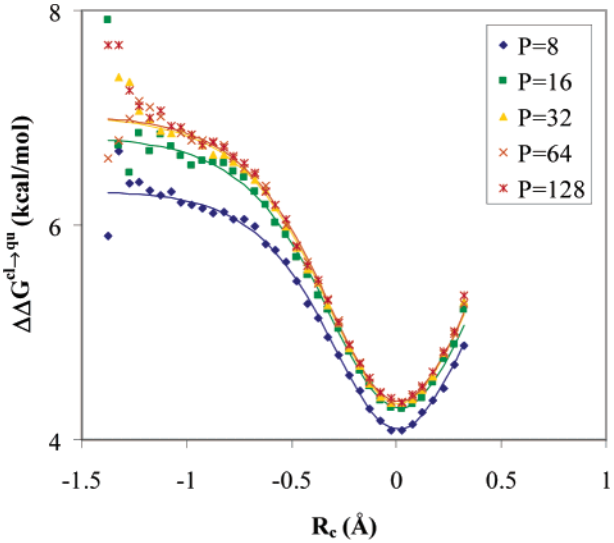


Figure 6. BQCP correction for the NH_4^+-NH_3 proton-transfer reaction in aqueous solution at different levels of path-integral discretization.

The extent of ring polymer beads required to obtain converged QM corrections was also investigated. Due to the great cost of PI simulations with a large number of beads, in the following test cases we used a scheme of 35 000 classical steps, each of which consists of 10 PI sampling steps. The results presented in Figure 6 show that the QM corrections to the computed classical free energy barrier increases with increasing number of beads. At the lowest level of discrete path description, using only 8 beads, we obtained a QM correction of 2.21 kcal/mol. Increasing the number to 16 improves the results at 2.50 kcal/mol, and 32 beads give a value of 2.64 at double the cost. Employing 64 beads yields identical quantization energy at 2.64 kcal/mol. A QM correction of 2.66 kcal/mol is obtained when doubling the number of beads to 128, thus indicating that a reasonable converged value is achieved with 32 beads. Thus, as was observed in our previous studies, using 32 beads seems to be a reasonable compromise between accuracy and cost.

A direct comparison of the results obtained herein with the results of Garcia-Viloca et al.[13] is not feasible due to the different methods employed. In the previous work only quantized vibrations were accounted for, while the current BQCP simulations account for both quantized vibrations and tunneling. In the work of Garcia-Viloca et al.,[13] a quantum correction of 2.0 kcal/mol was obtained, although the contribution of the mode corresponding to the reaction coordinate was not reported in that work. Thus, we performed instantaneous normal-mode analysis of the trajectories from the current simulations as well as tunneling calculations

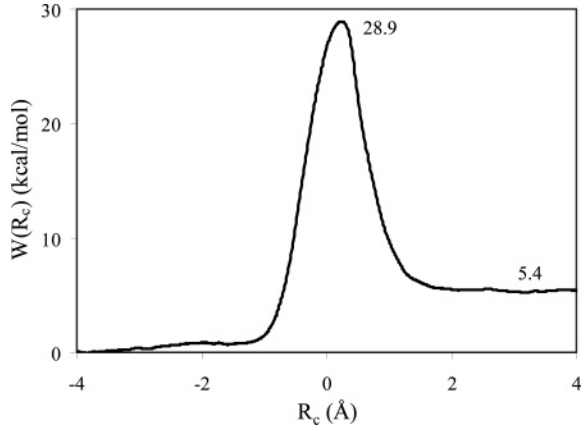


Figure 7. Classical potential of mean force for the deprotonation of nitroethane by an acetate ion in aqueous solution.

within the ensemble-averaged variational transition-state theory with multidimensional tunneling (EA-VTST/MT) approximation. The vibrational contribution to the quantized free energy amounts to 2.4 kcal/mol, in reasonable accord with ref 13 considering the small difference in the definition of the reaction coordinate. Tunneling does not seem to play a major role in the present system. The absence of tunneling is somewhat surprising considering the study of Truhlar et al.,[54] where tunneling was found to play a role in increasing the reaction rate. A reason for this difference may be the PM3 Hamiltonian used in that study, as opposed to AM1 which was utilized here. Indeed, the PM3 reaction barrier in solution was found to be in the range 13.3−15.9 kcal/mol,[54] compared to 6.1 kcal/mol obtained here with AM1.

B. Deprotonation of Nitroethane by Acetate Ion. The second reaction investigated is the ionization of nitroethane by an acetate ion in water (Figure 1B). This reaction is of particular interest because it is an analogue of the ionization of nitroethane by nitroalkane oxidase (NAO).[55−57] In this reaction, the active site residue Asp402 is the nucleophile that abstracts the α-proton of small nitroalkane substrates.[57] Moreover, the experimental reaction rate and kinetic isotope effects at several temperatures are available for this reaction.[55] The classical PMF yields a barrier of 28.90 kcal/mol (Figure 7), placing it slightly above the experimentally observed value of 24.8 kcal/mol.[55] Addition of nuclear quantum effects, which are dominated by the zero-point energy, lowers this classical value, as discussed below.

In the present test, a crucial question is the number of atoms to be quantized in the centroid PI simulation. In light of the conclusions obtained for reaction I, the BQCP simulations for reaction II used 32 beads and 10 PI steps per classical step, unless otherwise stated. The classical trajectories were saved every 25 steps.

Initially, only the transferring proton was quantized, yielding a QM correction value of 2.75 kcal/mol and a combined reaction barrier of 26.15 kcal/mol. A total of 27 000 classical configurations from 10 windows of umbrella sampling calculations were used to obtain the results. Comparison of results using different number of classical (centroid) configurations indicates that the results have converged. At the next level of quantization, we treated the three core transferring solute atoms as ring polymers. This allows for additional quantum effects to be included. We obtained a quantum correction of 3.03 kcal/mol, giving a reaction barrier of 25.87 kcal/mol in agreement with experiment.[45,55] A total of 38 000 classical steps were used to obtain the results. To better account for multidimensional nuclear QM effects, we additionally quantized the three neighboring heavy atoms (bonded to the transferring atoms). To obtain converged results, a total of 45 000 classical structures were employed, spanning 1.1 ns of MD simulations. Thus, we obtained a QM correction of 3.11 kcal/mol, placing the computed barrier of 25.79 kcal/mol in agreement with the experimental value. Although the difference is rather small, there is a gradual increase in the quantum correction as the number of quantized particle increases. Thus, to account for the QM correction to the barrier height of a classical PMF, it is important to include the donor and acceptor atoms in addition to the transferring light atom. This is in contrast to the ammonium ion-ammonia reaction where it seems to be sufficient only to quantize the transferring light particle. This difference is due to the greater rehybridization involved in the nitroethane-acetate ion reaction. Additionally, including the neighboring atoms has a small but noticeable effect on the QM correction to the barrier. However, inclusion of additional atoms impedes on the convergence of the free particle PI sampling, thus requiring additional sampling. The dependence of the QM correction on the number of classical configurations is illustrated in Figure 8 for the largest QM system. With a bin size of 0.001 Å, corresponding to ca. 20 configurations per bin, $\Delta\Delta G^{cl\rightarrow qu}$ fluctuates greatly. However, as the bin size increases, and thereby also the number of configurations per bin, the variance within each bin is reduced. When increasing the bin size from 0.05 to 0.1 Å no considerable change in $\Delta\Delta G^{cl\rightarrow qu}$ is observed, indicating that convergence with respect to classical configurations has been reached.

To verify that the use of 32 beads is reasonable for the current system, we also experimented with 16, 64, and 128 beads per quantized particle. To this end, we chose a quantized subsystem consisting of the transferring proton and the donor and acceptor atoms. Using 16 beads the QM correction was estimated as 2.86 kcal/mol, somewhat below the value obtained with 32 beads (3.03 kcal/mol). Increasing the number to 64 and 128 beads both yield a slightly larger value of 3.06 kcal/mol. Thus, employing 32 beads seems a reasonable choice in the case of reaction II as well.

Additionally, the second Cα hydrogen was added to the PI atom list, and although this increased the absolute QM correction value, it had a small effect on the barrier height.

To further test the BQCP method, we also studied the convergence of the computed KIE. In particular, we tested the importance of the number of quantized solute atoms, in obtaining reliable computed KIE. Quantizing only the transferring proton, the computed QM correction difference between proton and deuteron transfer is 1.06 kcal/mol, with a minimum in the QM correction curve at 0.1 Å. When combined with the classical PMF, this yields a computed KIE of 5.4 for a singly deuterated nitroethane. The total KIE is increased to 6.0 when secondary effects (1.10) are accounted for.[45] This may be compared to the experimental

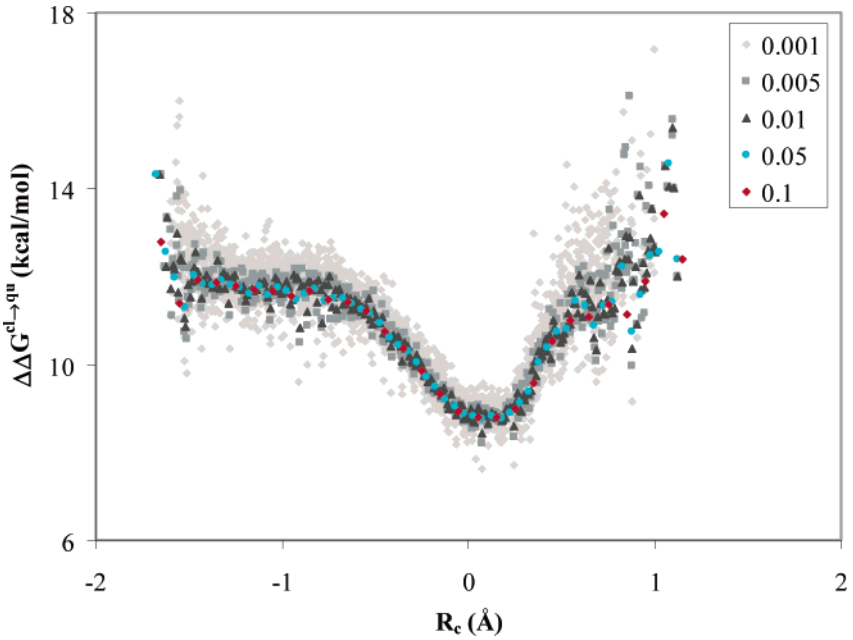


Figure 8. Computed average quantum corrections to the classical potential of mean force for the deprotonation of nitroethane by an acetate ion in aqueous solution, determined using different bin sizes in the quantized classical path averaging. Smaller bin sizes give relatively larger fluctuations, whereas a large bin width yields smooth results along the reaction path.

value of 7.8.[55] In free energy terms, the difference is about 0.1 kcal/mol. Increasing the number of beads to 64 and 128 did not narrow the difference between the experimental and the computed KIE (results not shown). Using 16 beads resulted in a considerable lower KIE of 4.3. Quantizing the donor and acceptor atoms, in addition to the transferring proton, using 32 beads resulted in similar QM correction difference between proton and deuteron transfer of 1.04 kcal/mol.

Thus, the current results indicate that approximately 32 beads are required to obtain reliable QM corrections to a classical PMF when using the QCP method. This is true also when computing KIE. Furthermore, to accurately compute QM corrections to a classical PMF, it is desirable to quantize the donor and acceptor atoms directly involved in the reaction, in addition to the transferring light particle. Similar findings have been observed in a previous study of the proton-transfer reaction of acetylacetone by Hinsen and Roux[53] and of malonaldehyde by Tuckerman and Marx.[52] Although we have not made a direct comparison, the efficiency of the present bisection sampling scheme should be similar to the staging approach described by Sprik et al.[21,26,58] Although the current study demonstrates that to obtain accurate KIE, it is sufficient to quantize the transferred particle alone, this is likely to be system dependent. In other systems possessing more extensive tunneling than what was observed for the current systems, multidimensional QM effects may be of importance.

Conclusions

The nuclear quantum mechanical effects in two aqueous solution proton-transfer reactions were investigated by a hybrid approach, combining QM/MM MD umbrella sampling simulations with the Quantized Classical Path (QCP) method described by Warshel and co-workers. The QCP method was augmented by the bisection algorithm (BQCP) to sample the free particle distribution and was shown to perform considerably better than when using the standard Metropolis method to sample the ring polymer chain. A sampling scheme was suggested that comprises a practical compromise between accuracy and computational cost. Additionally, different numbers of beads were tested, and the optimal choice for the systems studied was 32 beads. The conclusions found herein will be important for future studies of enzymatic reactions.

Acknowledgment. This work has been supported by the National Institutes of Health, and D.T.M. and M.G.-V. are Fulbright Scholars.

References

(1) Gao, J.; Truhlar, D. G. *Annu. Rev. Phys. Chem.* **2002**, *53*, 467−505.

(2) Garcia-Viloca, M.; Gao, J.; Karplus, M.; Truhlar, D. G. *Science* **2004**, *303*, 186−195.

(3) Benkovic, S. J.; Hammes-Schiffer, S. *Science* **2003**, *301*, 1196−1202.

(4) Kohen, A.; Limbach, H. H., Eds. *Isotope Effects in Chemistry and Biology*; Taylor & Francis Group, CRC Press: New York, 2005.

(5) Cha, Y.; Murray, C. J.; Klinman, J. P. *Science* **1989**, *243*, 1325−1330.

(6) Kohen, A.; Cannio, R.; Bartolucci, S.; Klinman, J. P. *Nature* **1999**, *399*, 496−499.

(7) Basran, J.; Sutcliffe, M. J.; Scrutton, N. S. *Biochemistry* **1999**, *38*, 3218−3222.

(8) Faulder, P. F.; Tresadern, G.; Chohan, K. K.; Scrutton, N. S.; Sutcliffe, M. J.; Hillier, I. H.; Burton, N. A. *J. Am. Chem. Soc.* **2001**, *123*, 8604−8605.

(9) Alhambra, C.; Corchado, J.; Sanchez, M. L.; Garcia-Viloca, M.; Gao, J.; Truhlar, D. G. *J. Phys. Chem. B* **2001**, *105*, 11326−11340.

(10) Billeter, S. R.; Webb, S. P.; Agarwal, P. K.; Iordanov, T.; Hammes-Schiffer, S. *J. Am. Chem. Soc.* **2001**, *123*, 11262−11272.

(11) Alhambra, C.; Luz Sanchez, M.; Corchado, J.; Gao, J.; Truhlar, D. G. *Chem. Phys. Lett.* **2001**, *347*, 512−518.

(12) Cui, Q.; Elstner, M.; Karplus, M. *J. Phys. Chem. B* **2002**, *106*, 2721−2740.

(13) Garcia-Viloca, M.; Alhambra, C.; Truhlar, D. G.; Gao, J. *J. Chem. Phys.* **2001**, *114*, 9953−9958.

(14) Tresadern, G.; Nunez, S.; Faulder, P. F.; Wang, H.; Hillier, I. H.; Burton, N. A. *Faraday Discuss.* **2002**, *122*, 223−242.

(15) Hwang, J. K.; Warshel, A. *J. Phys. Chem.* **1993**, *97*, 10053−10058.

(16) Hwang, J.-K.; Warshel, A. *J. Am. Chem. Soc.* **1996**, *118*, 11745−11751.

(17) Feierberg, I.; Luzhkov, V.; Aqvist, J. *J. Biol. Chem.* **2000**, *275*, 22657−22662.

(18) Major, D. T.; Gao, J. *J. Mol. Graphics Modell.* **2005**, *24*, 121−127.

(19) Ceperley, D. M. *Rev. Mod. Phys.* **1995**, *67*, 279−355.

(20) Ceperley, D. M.; Pollock, E. L. *Phys. Rev. Lett.* **1986**, *56*, 351−354.

(21) Sprik, M.; Klein, M. L.; Chandler, D. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1985**, *31*, 4234−4244.

(22) Keirstead, W. P.; Wilson, K. R.; Hynes, J. T. *J. Chem. Phys.* **1991**, *95*, 5256−5267.

(23) Voth, G. A.; Chandler, D.; Miller, W. H. *J. Chem. Phys.* **1989**, *91*, 7749−7760.

(24) Gillan, M. J. *J. Phys. Chem. A* **1987**, *20*, 3621.

(25) Chakrabarti, N.; Carrington, T., Jr.; Roux, B. *Chem. Phys. Lett.* **1998**, *293*, 209−220.

(26) Marx, D.; Tuckerman, M. E.; Martyna, G. J. *Comput. Phys. Comm.* **1999**, *118*, 166−184.

(27) Mielke, S. L.; Truhlar, D. G. *Chem. Phys. Lett.* **2003**, *378*, 317−322.

(28) Hwang, J. K.; Chu, Z. T.; Yadav, A.; Warshel, A. *J. Phys. Chem.* **1991**, *95*, 8445−8448.

(29) Hwang, K. Y.; Cho, C.-S.; Kim, S. S.; Sung, H.-C.; Yu, Y. G.; Cho, Y. *Nature Struct. Biol.* **1999**, *6*, 422−426.

(30) Olsson, M. H. M.; Siegbahn, P. E. M.; Warshel, A. *J. Am. Chem. Soc.* **2004**, *126*, 2820−2828.

(31) Feynman, R. P.; Hibbs, A. R. *Quantum Mechanics and Path Integrals*; McGraw-Hill: New York, 1965.

(32) Voth, G. A. *Adv. Chem. Phys.* **1996**, *93*, 135−218.

(33) Pollock, E. L.; Ceperley, D. M. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1984**, *30*, 2555−2568.

(34) Levy, P. *Compos. Math.* **1939**, *7*, 283.

(35) Press: W. H.; Flannery, B. P.; Teukolsky, S. A.; Vetterling, W. T. *Numerical Recipes*; University of Cambridge: New York, NY, 1992.

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

(37) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926−935.

(38) Gao, J. In *Rev. Comput. Chem.*; Lipkowitz, K. B., Boyd, D. B., Eds.; VCH: New York, 1995; Vol. 7, pp 119−185.

(39) Gao, J.; Thompson, M. A. *Combined Quantum Mechanical and Molecular Mechanical Methods*; American Chemical Society: Washington, DC, 1998; Vol. 712.

(40) Gao, J.; Xia, X. *Science* **1992**, *258*, 631−635.

(41) Bentzien, J.; Muller, R. P.; Florian, J.; Warshel, A. *J. Phys. Chem. B* **1998**, *102*, 2293−2301.

(42) Field, M. J.; Bash, P., A.; Karplus, M. *J. Comput. Chem.* **1990**, *11*, 700−733.

(43) Giese, T. J.; Sherer, E. C.; Cramer, C. J.; York, D. M. *J. Chem. Theory Comput.* **2005**, *1*, 1275−1285.

(44) Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Rassolov, V.; Pople, J. A. *J. Chem. Phys.* **1998**, *109*, 7764−7776.

(45) Major, D. T.; York, D. M.; Gao, J. *J. Am. Chem. Soc.* **2005**, *127*, 16374−5.

(46) Nam, K.; Gao, J.; York, D. M. *J. Chem. Theory Comput.* **2005**, *1*, 2−13.

(47) Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids*; Oxford University Press: Oxford, 1987.

(48) Valleau, J. P.; Torrie, G. M. In *Modern Theoretical Chemistry*; Berne, B. J., Ed.; Plenum: New York, 1977; Vol. 5, pp 169−194.

(49) Rajamani, R.; Naidoo, K.; Gao, J. *J. Comput. Chem.* **2003**, *24*, 1775−1781.

(50) Kumar, S.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A.; Rosenberg, J. M. *J. Comput. Chem.* **1992**, *13*, 1011.

(51) Gao, J. *Int. J. Quantum Chem.: Quantum Chem. Symp.* **1993**, *27*, 491−499.

(52) Tuckerman, M. E.; Marx, D. *Phys. Rev. Lett.* **2001**, *86*, 4946−4949.

(53) Hinsen, K.; Roux, B. *J. Chem. Phys.* **1997**, *106*, 3567−3577.

(54) Chuang, Y.-Y.; Cramer, C. J.; Truhlar, D. G. *Int. J. Quantum Chem.* **1998**, *70*, 887−896.

(55) Valley, M. P.; Fitzpatrick, P. F. *J. Am. Chem. Soc.* **2004**, *126*, 6244−6245.

(56) Fitzpatrick, P. F.; Orville, A. M.; Nagpal, A.; Valley, M. P. *Arch. Biochem. Biophys.* **2005**, *433*, 157−165.

(57) Valley, M. P.; Fitzpatrick, P. F. *J. Am. Chem. Soc.* **2003**, *125*, 8738−8739.

(58) Martyna, G. J.; Hughes, A.; Tuckerman, M. E. *J. Chem. Phys.* **1999**, *110*, 3275−3290.

CT050257T

JCTC Journal of Chemical Theory and Computation

Influence of Long-Range Electrostatic Treatments on the Folding of the N-Terminal H4 Histone Tail Peptide

Roberto D. Lins* and Ursula Röthlisberger

École Polytechnique Fédérale de Lausane, Institute of Chemical Sciences and Engineering, CH-1015 Lausanne, Switzerland

Received July 12, 2005

Abstract: A series of ca. 20-ns molecular dynamics simulation runs of the N-terminal H4 histone tail in its un- and tetraacetylated forms were performed using three different long-range electrostatic treatments namely, spherical-cutoff, reaction field, and particle mesh Ewald. Comparison of the dynamical properties of the peptide reveals that internal flexibility and sampling of the conformational space are heavily dependent on the chosen method. Among the three tested methods, the particle mesh Ewald treatment yields the least conformational variation and a structural stabilization tendency around the initially defined topological framework.

1. Introduction

Lattice summation methods are currently used as the standard long-range electrostatic treatment in explicit-solvent simulations using periodic boundary conditions. At least one out of a variety of implementations of the method (Ewald summation,[1] particle mesh Ewald (PME),[2] smooth particle mesh Ewald (SPME),[3] particle−particle particle mesh (P3M)[4]), originally developed for crystalline systems, has been implemented in the major biomolecular simulation softwares. Enforcing artificial periodicity by the means of lattice sum algorithms in inherent nonperiodic systems, such as explicitly solvated proteins and DNA, have been suggested to cause an unrealistic stabilization of the molecular system by reducing conformational sampling.[5−9] Recent studies conclude that although artificial periodicity induces a non-negligible energetic bias, it does not produce major structural perturbations in the solute.[10] However, the authors also state that the use of lattice sum methods overstabilize the secondary structure elements in high-temperature simulations.[10] In contrast, simulations of highly charged proteins[11] show marginal higher atomic positional fluctuations and deviations when P3M is compared to the reaction field (RF)[12] scheme. Most of the up-to-date conclusions are drawn based on relatively short time scale simulations (up to 3 ns) of stable protein or DNA structures. Therefore, to access the effect of three different electrostatic treatments, spherical group charge-based cutoff without switching function (SC),[13] RF and PME, on the structural variation of a conformationally rich peptide, a series of ca. 20 ns long explicit-solvent molecular dynamics (MD) simulations was performed. The 23-aa long amino-terminal H4 histone tail peptide was chosen for this purpose due to its small size and structural behavior change upon lysine acetylation. CD-spectra of the N-terminal H4 histone tails, in a 90% TFE (v/v) solution, show increasing helical content as a function of the number of acetylated lysines.[14,15] Due to its well-known α-helical stabilizer properties,[16−19] this solvent has been widely used to examine the helical propensity of peptides. The observed 25% helical propensity for the (fully) tetraacetylated form corresponds to ca. 5−6 residues.[15] This number is in good agreement with the length of the consensual helical region, residues 15−21, predicted by four independent secondary structure assignment methods.[15] It is worth noting that, generally, short-length peptides exhibit an inherent high flexibility when immersed in high-dielectric solvents. The chosen test-case makes no exception to this rule. However, major structural features of the system have been characterized, as discussed above. While the choice of a peptide of a well-defined structure would initially seem to be the ideal choice, it might also flatten out the foreseen differences this study aims at.

The present scope lies in a systematic evaluation of the impact of different commonly used electrostatic treatments

* Corresponding author phone: (509)375-2755; fax: (509)372-4720; e-mail: roberto.lins@pnl.gov. Present address: Pacific Northwest National Laboratory, Computational Biology and Bioinformatics, Richland, WA 99352.

10.1021/ct0501699 CCC: $33.50
Published on Web 02/01/2006

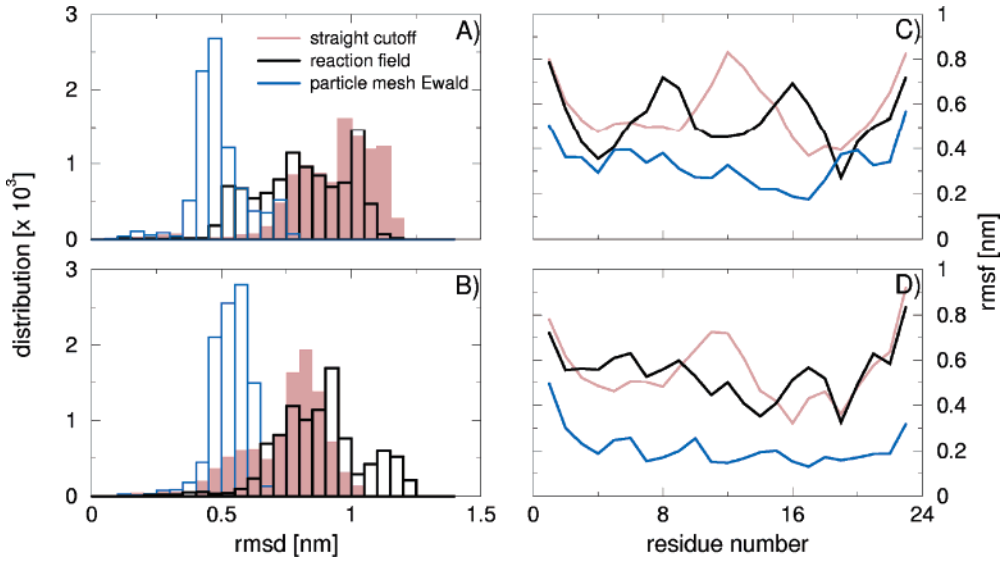


Figure 1. rmsd distributions for the backbone atoms of the H4 histone tail in its (A) non- and (B) tetraacetylated forms, and their corresponding rmsf (C and D, respectively), upon the three probed electrostatic treatments: SC (maroon), RF (black), and PME (blue). Analyses are displayed over the 0−18 ns window.

on conformational sampling of a biologically relevant system by the means of molecular dynamics simulations. Additionally, the results provide insights to the structure of the N-terminal H4 histone tails upon lysine acetylation.

2. Methods

Initial molecular dynamics simulation setups comprised the 23-residues N-terminal H4 histone tail peptide in a canonical α-helical conformation in its unacetylated (*nonac*) and tetraacetylated (*ac*) forms solvated in water. (The ability of these two systems to explore the conformational space was probed using three different methodologies to treat the long-range electrostatic interactions: spherical group charge-based cutoff without switching function (SC),[13] reaction field (RF),[12] and particle mesh Ewald (PME)[2]). The peptides measured a maximum of 3.6 nm in their longest axis and were solvated in a 7.0 × 7.0 × 7.0 nm SPC water[20] box, so to provide plenty of room between its periodic images. Counterions were added to both systems in order to keep their total charge equal to zero (0 *e*). The unacetylated system comprised 34 064 atoms (228 protein atoms, 9 Cl^- ions, and 11 275 water molecules), while the tetraacetylated system contained a total of 34 062 atoms (232 protein atoms, 5 Cl^- ions, and 11 279 water molecules). The systems were energy-minimized using 200 steepest decent steps. Equilibration was performed for 5 ps each at temperature intervals of 50, 100, 150, 200, 250, and 300 K, with velocity reassignment every 0.5 ps and a 2-fs time step. Production runs ranged from 18 to 25 ns in the NpT ensemble. The temperature was maintained at 300 K by a weak coupling to two independent heat baths with relaxation times of 0.1 ps for the solvent and the solute. The pressure of the system was kept at 1 bar by isotropic coordinate scaling with a relaxation time of 0.4 ps. SHAKE constraints[21] with a tolerance of 10^{-4} nm were applied to all bonds involving a hydrogen atom. A double twin-range cutoff of 0.8/1.4 nm was used when the long-range electrostatic interactions were treated by the cutoff-based methods (SC and RF). The short-range neighbor-list was updated every step, and the long-range one every 5 steps. A 1.0 nm cutoff was used as a short-range cutoff when PME was employed in the treatment of the long-range electrostatic contributions. All simulations were carried out using the GROMOS96 43A1 force field[22] within the Gromacs 3.2.1 package.[23] Coordinate frames were saved every 0.2 ps for analysis. The secondary structure content maps shown in the paper were performed via the DSSP program[24] implemented in the Gromacs program, version 3.2.1.[23]

3. Results and Discussion

The amount of conformational sampling was accessed via the root-mean-square deviation (rmsd) distribution for the backbone atoms over the 0−18 ns window (based on the shorter simulation) (Figure 1A,B). The PME treatment produced a narrower distribution in both the nonac and ac systems. Average rmsd is also relatively closer to its reference, i.e., a canonical α-helix in this case. The corresponding root-mean square fluctuation (rmsf) (Figure 1C,D) shows a reduced flexibility of 40−60% on average for the peptide treated by PME compared to the cutoff-based methods. The average overall fluctuation in the SC simulation is only ca. 3−7% higher than in the RF one. However, individual residue flexibilities and rmsd distributions indicate that these methods sample different parts of the phase-space for a given system at the nanosecond time scale.

The behavior of the secondary structure content of the peptides was analyzed as a function of the simulation time (Figure 2). The helical content in the nonac system is completely abolished within 1 and 4 ns when SC (Figure 2A) and RF methods (Figure 2B) are used. The PME treatment stabilizes the helical region spanning residues 14 to 22 in a 5-helix (*i*,*i*+5) configuration (Figure 2C). A higher helical content is observed for the acetylated form of the peptide regardless of the electrostatic treatment used. However, the actual value of helical content differs significantly. The use of SC and RF results in a α-helical propensity ca. 18% (in both cases) spanning residues 14 to 19 (Figure

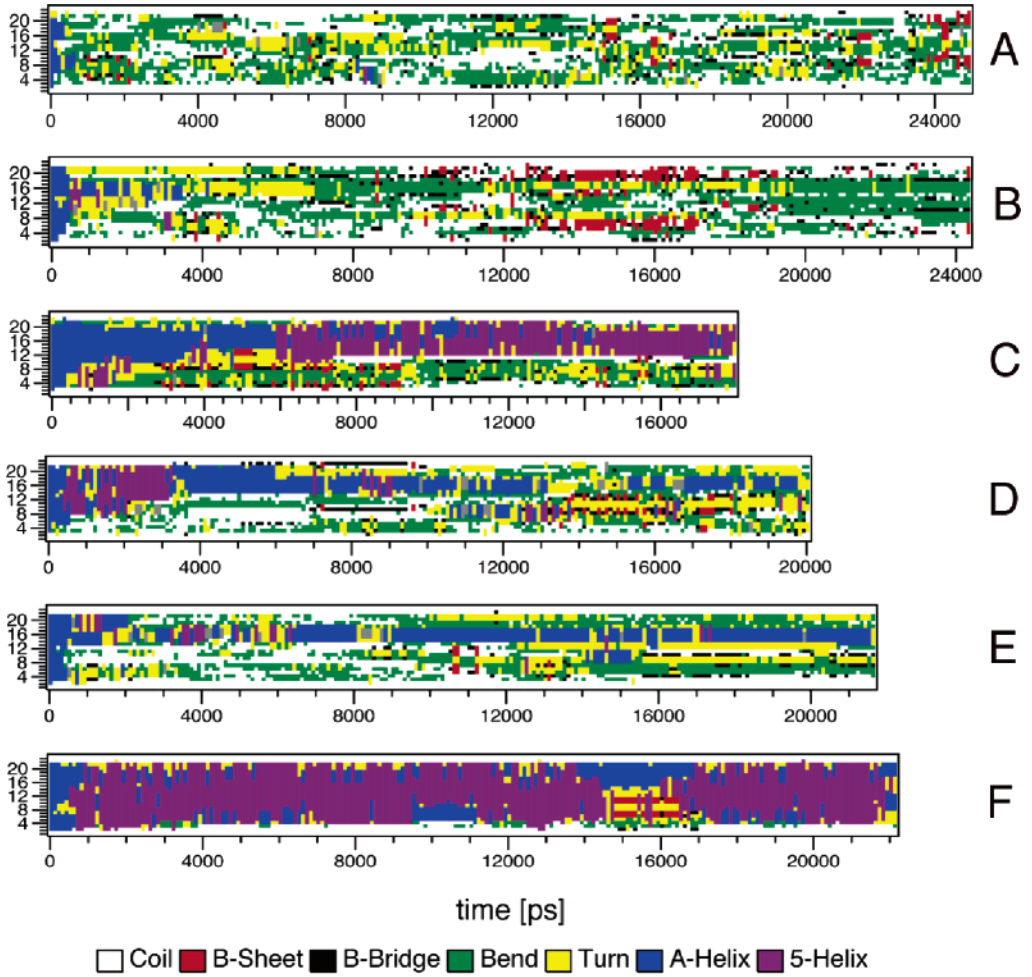


Figure 2. Secondary structure content for the N-terminal H4 histone tail in its unacetylated (nonac) and tetraacetylated (ac) forms as a function of the simulation time upon the three probed electrostatic treatments: (A) nonac/SC, (B) nonac/RF, (C) nonac/PME, (D) ac/SC, (E) ac/RF, and (F) ac/PME. (Secondary elements are defined according to the DSSP program.[20] The color-coded chart at the bottom is provided for content identification.)

2D,E). The helical occupancy is about 9% higher in the RF simulation. This increased stability may have its roots in the better energy conservation scheme used by this method over the SC one. PME treatment of the acetylated peptide is characterized by a mix of 5- and α-helix structure (except for ca. 10% of the time (∼13.6−17 ns) (Figure 2F)). These findings indicate that either (i) the use of the cutoff-based methods (SC and RF, here) would produce an enhanced sampling due to energy conservation issues, or (ii) the intrinsic artifacts arisen from PME would cause a reduction in the conformational sampling. The inability to conserve energy is a well-known limitation of the SC method. However, few variations such as the RF or the more recently developed force-shifted spherical cutoff methods[21] have proven to keep this issue to a minimum[12,25] and produce conformational samplings and distributions in excellent agreement with experimental (NMR, CD, ORD, X-ray crystallography) data.[25−32] In addition, based on the CD-spectra and the secondary content prediction data[14,15] the PME treatment seems to overestimate the total helical content in these runs. It results in over 40% and 70% helical content for the nonac and ac simulations, respectively.

In a recent study, Monticelli and Colombo have compared the ability of PME- and SC-simulations of the β3 peptide to reproduce experimental NOE restraints.[33] Based on 500-ns simulations, the use of PME allowed a better description of the experimental NOE restraints and secondary structure content than the SC simulations. A comparison of the different electrostatic methodologies was not the primary intention of this study, and no systematic assessment has been made in this direction. However, their findings suggested an overestimation of the secondary structure content, reduced flexibility, and stabilization tendency around the initial conformation when PME is used.[33] At the same time, the SC simulations have failed to provide an accurate description of the NOE restraints apparently due to a high flexibility.[33] No alternative methodology has been tested or proposed in order to overcome the problem. To evaluate this apparent overstabilization tendency from the lattice-sum method, a frame without any helical content was randomly extracted from the ac/SC simulation (at 14 280 ps) and switched to the PME treatment. A secondary structure map as a function of time of this new 23+ ns MD simulation is shown in Figure 3. The results revealed indeed an entrapment of the peptide around the newly provided topological framework.

While reverse folding is not expected in a 24+ ns MD simulation, the immutability of the completely distinct

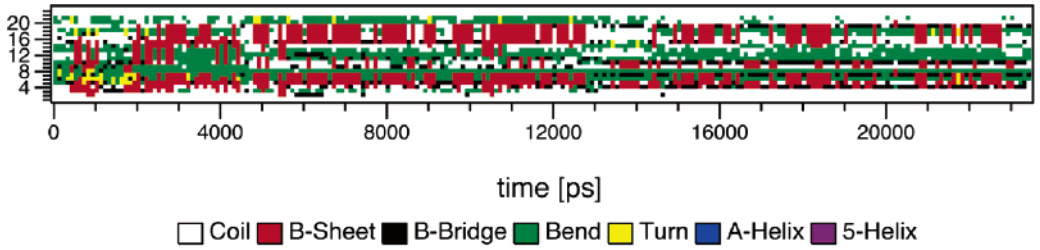


Figure 3. Secondary structure content for the tetraacetylated (ac) form of the N-terminal H4 histone tail as a function of the simulation time. (Secondary elements are defined according to the DSSP program.[24] The color-coded chart at the bottom is provided for content identification.)

secondary elements in the simulations of the same peptide displayed in Figures 2F and 3 is remarkable in such time scale, especially if compared with the cutoff-based runs shown in Figure 2.

4. Conclusion

The limitations of the SC method in molecular dynamics simulations are well-established ones and known to lead to severe problems in the energy conservation scheme and unrealistic structural distortions.[34−39] A number of studies, using different force fields, have shown that correction methods applied outside of the cutoff sphere, such as the continuum-based reaction field and switching methods, may represent a satisfactory compromise between accuracy and the computational costs associated with the different Ewald summation methods.[10,11,25,28,31] This conclusion is partly due to the marginal difference observed in sampling and thermodynamical properties when comparing simulations using long-range correction techniques with PME-treated ones.[10,11,28,31] However, these independent studies involved either relatively short simulation times (up to 3 ns) of highly stable proteins and DNA[10,11,28] or enhanced sampling simulations[31] (high temperatures/replica-exchange algorithms). While the latter case is undeniably more efficient, it may flatten out the effect of the long-range electrostatic treatments on conformational sampling at room temperatures. Nevertheless, the exchange of replicas acceptance ratio is reported as slightly smaller for the peptides in the PME simulation in comparison to the RF ones.[31] It might suggest to some extent a higher overall conformational hindrance of the system treated by PME. The outcome of the simulations presented here, however, shows clearly that (i) the sampling of the conformational phase-space in a typical molecular dynamics simulation can be heavily influenced by the choice of the long-range electrostatic treatment and, consequently, (ii) that RF may not always be used as a cost-efficient alternative to PME. It is worth noting that these discrepancies in sampling are likely to be less pronounced in larger and conformationally stabler molecular systems. However, tests at the present or, ideally, longer simulation lengths would be required in order to confirm such an assumption.

In summary, the present simulations consistently show that the acetylated form of the H4 histone tail contains a higher helical content regardless of the electrostatic treatment used. This helical structure occurs toward the C-terminal region of the peptide in agreement with experimental and secondary structure prediction data.[14,15] The use of the SC, RF, and PME methods, within the same time scale and at room temperature, seems to yield the exploration of different regions of the energy surface for a given system. The peptides treated by the PME scheme show the least conformational variation throughout the dynamics and a non-negligible overstabilization of the secondary structure. Sampling is reduced when compared to the SC and RF methods and, to some extent, restricted to the initial structural topology defined in the setup. Therefore, the current findings suggest that the use of cutoffs along with proper correction methods to the outside-cutoff sphere may be better suited for the sampling of the conformational space of highly flexible naturally inhomogeneous systems, such as the study of protein and peptide folding. However, substantiation of the current observation is not an easy task since it would require the similar assessment of systems that are simultaneously internally flexible and structurally well characterized. While such entities are far from abundant, in the past few years the groups of Seebach and van Gunsteren have combined NMR experiments and MD simulations to characterize the structure and dynamics of several fast-reversible-folding β-peptides.[26,27,40,41] Given the impressive level of agreement obtained between theory and experimental data, even when looking at different temperatures, it places these non-natural peptides as a potential test-case system for this type of study. Being currently limited to the nano/microsecond time scale the understanding of the influence of different electrostatic methods on the dynamics of biomolecules becomes imperative. The problem deepens if the effect of temperature and/or pressure is added, which is the case in multiple-replica MD simulations. Efforts are currently being made in this direction and shall be reported soon.

Acknowledgment. R.D.L. would like to thank Dr. Thereza Soares, Dr. Philippe Hünenberger, and Dr. Maurício Coutinho-Neto for insightful discussions.

References

(1) Ewald, P. P. *Ann. Phys.* **1921**, *64*, 253−287.

(2) Darden, T. A.; York, D.; Pedersen, L. G. *J. Chem. Phys.* **1993**, *98*, 10089−10092.

(3) Essman, U.; Perera, L.; Berkowitz, M. L.; Darden, T. A.; Lee, H.; Pedersen, L. G. *J. Chem. Phys.* **1995**, *103*, 8577−8593.

(4) Hockney, R. W.; Eastwood, J. W. *Computer simulation using particles*; McGraw-Hill: New York, 1981.

(5) Brooks, C. L. *Curr. Opin. Struct. Biol.* **1995**, *5*, 211−215.

(6) LouiseMay, S.; Auffinger, P.; Westhof E. *Curr. Opin. Struct. Biol*. **1996**, *6*, 289−298.

(7) Auffinger, P.; LouiseMay, S.; Westhof, E. *J. Am. Chem. Soc.* **1996**, *118*, 1181−1189.

(8) De Bakker, P. I. W.; Hünenberger, P. H.; McCammon, J. A. *J. Mol. Biol.* **1999**, *285*, 1811−1830.

(9) Hünenberger, P. H.; McCammon, J. A. *Biophys. Chem.* **1999**, *78*, 69−88.

(10) Kastenholz, M. A.; Hüneneberger, P. H. *J. Phys. Chem. B* **2004**, *108*, 774−788.

(11) Gargallo R.; Hünenberger P. H.; Aviles, F. X.; Oliva B. *Protein Sci*. **2003**, *12*, 2161−2172.

(12) Tironi I. G.; Sperb R.; Smith P. E.; van Gunsteren W. F. *J. Chem. Phys*. **1995**, *102*, 5451−5459.

(13) Allen M. P.; Tildesley D. J. *Computer Simulation of Liquids*; Clarendon Press: Oxford, 1987.

(14) Cary, P. D.; Crane-Robinson, C.; Bradbury, E. M.; Dixon, G. H. *Eur. J. Biochem.* **1982**, *127*, 137−143.

(15) Wang, X.; Moore, S.; Laszckak, M.; Ausió, J. *J. Biol. Chem.* **2000**, *275*, 35013−35020.

(16) Nelson, J. W.; Kallenbach, N. R. *Biochemistry* **1989**, *28*, 5256−5261.

(17) Lehrman, S. R.; Tuls, J. L.; Lund, M. *Biochemistry* **1990**, *29*, 5590−5596.

(18) Segawa, S.; Fukuno, T.; Fujiwara, K.; Noda, Y. *Biopolymers* **1991**, *31*, 497−509.

(19) Sonnichsen, F. D.; van Eyk, J. E.; Hodges, R. S.; Skyes, B. D. *Biochemistry* **1992**, *31*, 8790−8798.

(20) Berendsen, H. J. C.; Potsma, J. P. M.; van Gunsteren, W. F.; Hermans, J. In *Intermolecular Forces*; Pullman, B., Ed.; Reidel: Dordrecht, p 331.

(21) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977** *23*, 327−341.

(22) Daura, X.; Mark, A. E.; van Gunsteren, W. F. *J. Comput. Chem.* **1998**, *19*, 535−547.

(23) Lindahl, E.; Hess, B.; van der Spoel, D. *J. Mol. Mod.* **2001**, *7*, 306−317.

(24) Kabsh, W.; Sander, C. *Biopolymers* **1983**, *22*, 2577−2637.

(25) Beck, D. A.; Armen R. S.; Daggett V. *Biochemistry* **2005**, *44*, 609−616.

(26) Daura, X.; Gademann, K.; Jaun, B.; Seebach, D.; van Gunsteren, W. F. *Angew. Chem. Int. Ed.* **1999**, *38*, 238−240.

(27) Daura, X.; van Gunsteren, W. F.; Mark, A. E. *Proteins* **1999**, *15*, 269−280.

(28) Walser, R.; Hünenberger, P. H.; van Gunsteren, W. F. *Proteins* **2001**, *44*, 509−519.

(29) Chandrasekar, I.; Kastenholtz, M.; Lins, R. D.; Oostenbrink, C.; Schuler, L. D.; Tieleman, P. D.; van Gunsteren, W. F. *Eur. Biophys. J.* **2003**, *32*, 67−77.

(30) Soares, T. A.; Hünenberger, P. H.; Kastenholz, M.; Krautler, V.; Lenz, T.; Lins, R. D.; Oostenbrink, C.; van Gunsteren, W. F. *J. Comput. Chem*. **2005**, *26*, 725−737.

(31) Baumketner, A.; Shea, J.-E. *J. Phys. Chem. B* **2005**, *109*, 21322−21328.

(32) Lins R. D.; Hünenberger P. H. *J. Comput. Chem.* **2005**, *26*, 1400−1412.

(33) Monticelli, L.; Colombo, G. *Theor. Chem. Acc.* **2004**, *112*, 145−157.

(34) Loncharich, R. J.; Brooks, B. R. *Proteins* **1989**, *6*, 32−45.

(35) Cheetham, T. E., III.; Brooks, B. R. *Theor. Chem. Acc.* **1998**, *99*, 279−288.

(36) Smith, P. E.; Pettitt, B. M. *J. Chem. Phys.* **1991**, *95*, 8430−8441.

(37) Schreiber, H.; Steinhauser, O. *Biochemistry* **1992**, *31*, 5856−5860.

(38) York D. M.; Wlodawer A.; Pedersen L. G.; Darden T. A. *Proc. Natl. Acad. Sci. U.S.A.* **1994**, *91*, 8715−8718.

(39) Cheetham, T. E., III.; Miller, J. L.; Fox, T.; Darden, T. A.; Kollman, P. A. *J. Am. Chem. Soc.* **1995**, *117*, 4193−4194.

(40) Daura, X.; Jaun, B.; Seebach, D.; van Gunsteren W. F.; Mark, A. E. *J. Mol. Biol*. **1998**, *280*, 925−932.

(41) Daura, X.; Gademan, K.; Schaefer, H.; Jaun, B.; Seebach, D.; van Gunsteren W. F. *J. Am. Chem. Soc*. **2001**, *123*, 2393−2404.

CT0501699

JCTC Journal of Chemical Theory and Computation

Essential Dynamics: A Tool for Efficient Trajectory Compression and Management

Tim Meyer,[†,▼] Carles Ferrer-Costa,[†,▼] Alberto Pérez,[†] Manuel Rueda,[†] Axel Bidon-Chanal,[‡] F. Javier Luque,[‡] Charles. A. Laughton,[§] and Modesto Orozco*[,†,||,⊥,#]

Institut de Recerca Biomèdica Barcelona, Parc Científic de Barcelona, Josep Samitier 1-5, Barcelona 08028, Spain, Departament de Fisicoquímica, Facultat de Farmacia, Universitat de Barcelona, Avgda Diagonal 643, Barcelona 08028, Spain, School of Pharmacy and Centre for Biomolecular Sciences, University of Nottingham, Nottingham NG7 2RD, U.K., Departament de Bioquímica i Biologia Molecular, Diagonal 645, Barcelona 08028, Spain, Computational Biology Program, Barcelona Supercomputing Center, Jordi Girona 31, Edifici Torre Girona, Barcelona 08028, Spain, and Bioinformatics Structural Node, Instituto Nacional de Bioinformática, Josep Samitier 1-5, Barcelona 08028, Spain

Received November 21, 2005

Abstract: We present a simple method for compression and management of very large molecular dynamics trajectories. The approach is based on the projection of the Cartesian snapshots collected along the trajectory into an orthogonal space defined by the eigenvectors obtained by diagonalization of the covariance matrix. The transformation is mathematically exact when the number of eigenvectors equals $3N-6$ (N being the number of atoms), and in practice very accurate even when the number of eigenvectors is much smaller, permitting a dramatic reduction in the size of trajectory files. In addition, we have examined the ability of the method, when combined with interpolation, to recover dense samplings (snapshots collected at a high frequency) from more sparse (lower frequency) data as a method for further data compression. Finally, we have investigated the possibility of using the approach when extrapolating the behavior of the system to times longer than the original simulation period. Overall our results suggest that the method is an attractive alternative to current approaches for including dynamic information in static structure files such as those deposited in the Protein Data Bank.

Introduction

Recent advances in algorithms, force-fields, and computer power have greatly promoted the use of molecular dynamics (MD) simulations to gain deeper insight into the structural and dynamical behavior of biomolecules. MD is becoming a standard tool even for experimental groups, and trajectories are being collected for larger systems and for longer simulation times. A search of the *pubmed* server using the keyword "molecular dynamics" found 1608 entries for the period 1992−1994 and 6865 citations for 2002−2004. In addition, while 5 years ago *state-of-the-art* MD typically provided trajectories that covered around 10 ns for biomolecular systems of the size of 100-residue proteins or 12-mer DNAs, today the length of such simulations has increased by nearly 1 order of magnitude, and some groups are turning their attention to far larger systems such as the nucleosome or the ribosome.[1,2] The net result of all this

* Corresponding author e-mail: modesto@mmb.pcb.ub.es.
† Institut de Recerca Biomèdica Barcelona.
‡ Universitat de Barcelona.
§ University of Nottingham.
|| Departament de Bioquímica i Biologia Molecular.
⊥ Barcelona Supercomputing Center.
Instituto Nacional de Bioinformática.
▼ These authors contributed equally to this work.

10.1021/ct050285b CCC: $33.50
Published on Web 02/14/2006

activity is a huge increase in the quantity and quality of available MD data, and how these data can be efficiently stored and retrieved are becoming issues of concern.

The trajectory collected in a MD run consists of a very large file (or a sequence of smaller ones) containing a series of 'snapshots'—the coordinates of the system—over the simulation time. The integration algorithm provides the coordinates of all the atoms in the system every 1−2 fs, but data are output to file much less frequently (typically every 1 ps). Despite this, long trajectories of large systems generate huge data files (many gigabytes in size) that can place a severe burden on disk storage and data transfer systems, because the process of data analysis (which nowadays will usually take much longer than the MD simulation that generated the data) will typically require frequent and high-speed access to the data, very often from remote locations. Additionally, it is increasingly the case that data generated by one research group for one purpose is seen as potentially valuable to another group for another purpose, so questions of enabling the efficient archival and remote retrieval of the data become important. Examples of projects that are facing this issue include the ABC-database (http://max.chem.wesleyan.edu/), the BiosimGrid (www.biosimgrid.org) project, and the MODEL (htpp://mmb.pcb.ub.es/MODEL) project, which involve (i) generating and managing hundreds of very large trajectories for different systems (our group generates nearly 1Tb of trajectory data every month through the MODEL project) and (ii) processing, analyzing, and making available to the scientific community both the analysis and the 'raw' data itself.

In this paper we will present a method that exploits the concept of essential dynamics for the compression and management of large MD trajectories. The method allows a dramatic reduction in the size of the files with no significant loss of quality in the results. Furthermore, the reduction of noise implicit to the use of the method helps in the interpretation of the essential features of the trajectory. The algorithms presented here have been tested using a series of MD trajectories taken from our MODEL database as well as with a very long trajectory of a 28-mer DNA duplex.

Basic Approaches

Essential dynamics (ED) is a very powerful analysis technique[3−6] which exploits principal component analysis to identify the nature and relative importance of the essential deformation modes of a macromolecule from MD samplings. Accordingly, the original Cartesian covariance matrix which contains the atomic positional fluctuations in all 3 coordinate axis about the average structure is diagonalized to obtain a set of eigenvectors and eigenvalues. The eigenvectors describe the nature of deformation movements in Cartesian space, whereas the eigenvalues represent the amount of variance explained by each movement. The eigenvectors define a complete and orthogonal basis set, and accordingly any snapshot in the trajectory can be exactly reproduced in this new $3N-6$ basis set (N is the number of atoms in the system; see eq 1)

$$\{R\}_{x,y,z} \rightarrow \{P\}_{\nu} \tag{1}$$

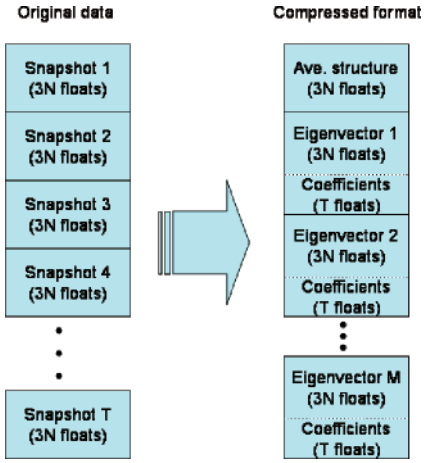


Figure 1. Schematic representation of the data structure of original and compressed trajectories.

where R stands for the original Cartesian (x,y,z) coordinates and P stands for the projections in the $3N-6$ eigenvectors (υ), which are defined to maximize the amount of variance explained in a descending order. Of course, the reverse process is also possible: the original data can be recovered by back-projection from the eigenvectors space to the Cartesian one.

For proteins and nucleic acids the number of important eigenvectors (i.e., those needed to explain 95−99% of the total variance) is much smaller than $3N-6$. If just M eigenvectors describe, say, 95% of the total variance (see below), then projections of the original Cartesian coordinates along the set of important eigenvectors contain nearly all of the original information in a much more compressed way, and it is still possible to regenerate Cartesian coordinate data ($\{R_j\}^n_{x,y,z}$) by back projection (eq 2), though the reconstituted coordinates are no longer exact.

$$\{R_j\}_{x,y,z} \rightarrow \{P_j\}^M_{\nu} \rightarrow \{R_j\}^n_{x,y,z} \tag{2}$$

The opportunities for data compression are obvious. For a set of T snapshots ($T > 3N-6$), the original trajectory file will contain $3NT$ coordinates. This will be transformed (see Figure 1) into a set of M eigenvectors, each of size $3N$, plus M sets of T coefficients—total $M(3N+T)$. For a typical current MD trajectory, reasonable values of N, M, and T would be 500, 50, and 2000, respectively. This would translate into compression of the data to 5.8% of its original size. The question then is what is the cost of this compression—i.e., what is the error between $\{R_j\}^n_{x,y,z}$ and $\{R_j\}_{x,y,z}$.

Possibilities for further data compression also exist. By the quasi-harmonic approximation, the modes of deformation (eigenvectors) associated with the largest eigenvectors are expected to show the lowest frequencies of motion. If the coefficients associated with these modes vary slowly with time (compared to the original snapshot sampling rate), then it should be possible to reduce the M sets of T coefficients to a more sparsely sampled set and regenerate intermediate values by a process of interpolation. Again, the question we seek to address here is to what extent this procedure is useable with 'real life' data.

As a partial aside, we also investigate the utility of this process for data extrapolation, rather than interpolation. We examine to what extent a set of *M* eigenvectors, chosen to be able to capture the dynamic behavior of a system to within a defined tolerance during the time period *T*, are able to continue to represent the system for times beyond *T*. This is not a new idea, forming as it does the basis of the approach of Essential Dynamics,[3−6] but here we provide a detailed analysis based on a wide range of representative systems, of the reliability of this approach.

Practical Derivation of Projections

Covariance matrices were created from equally spaced snapshots collected during long MD trajectories (see above). Following our previous studies,[7] unless specifically noticed, at least $3N-6$ snapshots were collected for each system. Time spacing for data collection ranged from 1 to 10 ps. Once the covariance matrix was defined, eigenvalues were computed using standard algebraic procedures which avoid memory-costly inversion procedures. The percentage of total variance explained by each essential movement is determined according to eq 3, where λ is the set of eigenvalues (in the same distance[2] units in which the covariance matrix is created) and *N* is the number of atoms in the macromolecule. We then determined the minimum number of eigenvectors needed to account for a given amount of variance (generally 95% or 99%), defining an "important space" of *M* eigenvectors which represent the main global movements.

$$\%\ \mathrm{var}_i = 100\frac{\lambda_i}{\sum_{i=1}^{3N-6}\lambda_i} \quad (3)$$

Following *Ptraj* implementation in the AMBER suite of programs,[8] we first derive the eigenvalues using Pal, Walker, and Kahan method,[9,10] whose computational cost scales with the square of the number of atoms. The Arnoldi-Lanczos[9,10] method is then used to find pairs of eigenvectors/eigenvalues in the reduced space (dimension *M*) corresponding to a given amount of variance (determined from the eigenvalues). This latter method is more efficient than the PWK one when a small number of eigenpairs needs to be determined. Finally, the original Cartesian coordinates are projected using the reduced space of eigenvectors (eq 3), which is not strictly exact since $M \ll 3N-6$.

Inter- and Extrapolation of Trajectories

Another goal of this study is to explore the possible use of the preceding procedure to interpolate trajectories, i.e., to estimate a trajectory sampled at a time interval t' from one originally collected with a time interval t and t' and $t' < t$ (eq 4)

$$\{R_j\}_{x,y,z}(t) \rightarrow \{P_j\}_{\nu}^{M}(t) \rightarrow \{P_j\}_{\nu}^{M}(t') \rightarrow \{R_j\}_{x,y,z}^{n}(t') \quad (4)$$

where *t* stands for the time used for storage of the original data, which is used to derive eigenvectors and projections, and t' is the new time spacing ($t'<t$) used to build up the new trajectory.

To this end, we explored the goodness of a simple linear interpolation scheme where a Gaussian noise (Θ) may be introduced to include some stochastic nature in the trajectories (eq 5). The use of the Gaussian noise (always set to define a standard deviation around 10%) helps to reduce an excessive correlation between the interpolated and the original points

$$\{P_j\}_{\nu}^{M}(tt + tt') = \{P_j\}_{\nu}^{M}(tt) + tt'\frac{\{P_j\}_{\nu}^{M}(tt + t) - \{P_j\}_{\nu}^{M}(tt)}{t} + \Theta(tt + tt') \quad (5)$$

where *tt* and *tt*+*t* are trajectory times at which trajectory points were originally collected and *tt*+*tt*′ stands for new times at which the trajectory is interpolated under the constraint that *tt*+*tt*′ pertains to the interval from *tt* to *tt*+*t*.

As discussed above, the eigenvectors obtained from a portion of a trajectory can be used to extrapolate the behavior of the system forward in time, allowing the rapid generation of very long pseudoharmonic trajectories. We explored the validity of this extrapolation scheme by projecting the Cartesian coordinates of a portion of trajectory $t \rightarrow t+\Delta t$ into the set of important eigenvectors obtained from a previous portion of the same trajectory (for example $t\text{-}\Delta t \rightarrow t$). The back-projection procedure generates a new set of Cartesian coordinates, which are then compared with the original ones and with those obtained when the process is repeated using eigenvectors obtained from the same portion of the trajectory (i.e., the period $t \rightarrow t+\Delta t$).

Simulation Details

The compression procedure was first examined using a long (70 ns) trajectory of the DNA duplex d(AAGCATTTTCACG-CATGAGTGCACAGAA). The simulation system contains a total of nearly 82 000 atoms (including water and counterions) and constitutes, to our knowledge, the longest DNA fragment ever simulated over this time scale in explicit solvent. It is therefore an excellent example with which to illustrate the possibilities of the compression procedure and to gain insight into the accuracy of the interpolation and extrapolation schemes outlined above.

The performance of the approach presented here was also examined using 10 ns trajectories of a small set of proteins, which a priori should have more complex dynamics than nucleic acids.[6,11] These were taken from our MODEL database (http://mmb.pcb.ub.es/MODEL). The selected set (PDB entries 1ark, 1cei, 1sro, 2gb1, 3ci2, and 4icb) contains examples of all-α, α+β, and all-β small globular proteins. Finally, we analyzed a long MD simulation (100 ns) of a medium-size protein (PDB entry 1idr) to evaluate the behavior of the method when dealing with long trajectories.

Simulations were performed in all cases using the AMBER parm98[12] force field and the TIP3P[13] model for water and suitable equilibration protocols (from 0.2 to 1 ns). All trajectories were performed in the isothermic-isobaric ensemble (298 K and 1 atm) using Particle Mesh Ewald[14] and truncated octahedral periodic boundary conditions. For the DNA simulation an integration time step of 1 fs was used in conjunction with SHAKE applied to bonds involving

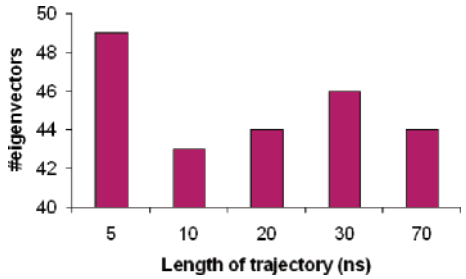


Figure 2. Number of eigenvectors needed to represent 95% of the variance in the trajectory of d(AAGCATTTTCACGCATGAGTGCACAGAA)$_2$ for different simulation times.

hydrogens,[15] while protein simulations were performed using a 2 fs integration time step and SHAKE for all covalent bonds. Simulations were performed using AMBER8.0[8] and NAMD2.5[16] programs on the MareNostrum Cluster (PowerPC64/Myrinet) at the Barcelona Supercomputing Center (details: http://www.bsc.es). All simulations were inspected to verify the lack of equilibration artifacts and that the structural parameters are similar to those determined experimentally. The analysis was performed using a modified version of the *Ptraj* module of AMBER8.0[8] and *in-house* software.

Results and Discussion

DNA Simulations. Our previous studies[6,7,18] have shown that DNA has a quite simple dynamical behavior, which can be represented to a high degree of accuracy by a limited number of eigenvectors. Thus, though the 28-mer duplex considered here has around 1600 atoms, only 44 (220) essential modes are able to capture 95% (99%) of the variance in the 70 ns trajectory of the duplex. These numbers remain quite constant if shorter simulation times are considered, thus revealing that the complexity of conformational space sampled does not increase with the length of the simulation time (see Figure 2).

The average all-atom RMSd between the MD-averaged structure and the 7000 collected snapshots is around 3.0 ± 0.8 Å, with the largest deviations being around 6.5 Å (see Figure 3). When the projection→back-projection procedure is performed using only the first eigenvector (which explains 29% of variance), the RMSd is reduced to 2.5 ± 0.7 Å, and the largest RMSd is close to 6 Å. When the importance space is expanded to the first 5, 44, and 220 eigenvectors (76%, 95%, and 99% of the variance, respectively), the average RMSd is reduced to 1.5 ± 0.2 Å, 0.68 ± 0.06 Å, and 0.30 ± 0.02 Å, and the largest RMSd is below 2 Å when 5 eigenvectors are used and very close to the average error in the other two cases (see Figure 3). Such small errors are impossible to obtain by just taking "representative structures" obtained from clustering analysis of the bidimensional RMSd space. Not surprisingly, the coordinates generated with the compression procedure lead to helical parameters similar to the original ones for the 95% and 99% cutoff levels (see Table 1). The similarity is maintained when the helical

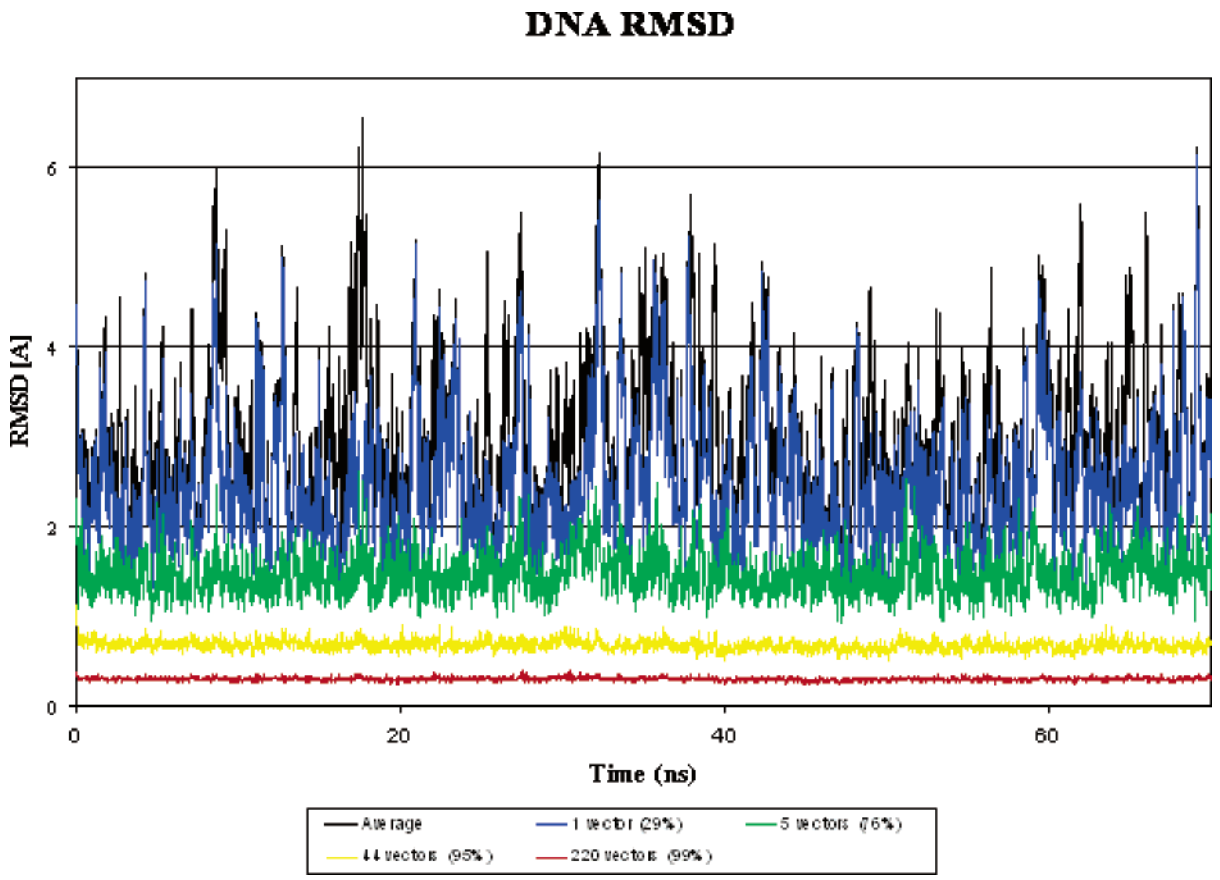


Figure 3. RMSd (in Å) between the real DNA-trajectory and coordinates generated using the projection→back-projection procedure with a different number of eigenvectors (1: blue; 5: green; 95% variance: yellow; 99% variance: red). The results obtained when no eigenvectors are used (average structure) are displayed, for reference, in black.

Table 1. Average Helical Parameters (with Standard Deviations) Associated with the 70 ns Trajectory of DNA Studied Here

helical parameter	original	99% cutoff	95% cutoff
shift	-0.05 ± 0.1	-0.05 ± 0.1	-0.05 ± 0.1
slide	-0.42 ± 0.2	-0.42 ± 0.2	-0.42 ± 0.1
rise	3.34 ± 0.03	3.34 ± 0.03	3.36 ± 0.03
tilt	-0.22 ± 0.6	-0.21 ± 0.6	-0.18 ± 0.2
roll	3.49 ± 1.0	3.49 ± 0.9	3.54 ± 0.9
twist	32.3 ± 0.6	32.3 ± 0.6	32.3 ± 0.6

analysis is performed at the base pair level (see Table S1 in Supporting Information).

As expected, the neglect of fast intramolecular vibrations in the projection→back-projection process generates some deviations of bond lengths and angles from the optimum values and eventually to some incorrect van der Waals contacts. However, these alterations do not affect key intramolecular interactions such as stacking or hydrogen-bond (see Table S2 in Supporting Information). In any case, these artifacts can be easily eliminated by a few cycles of geometry optimization without any significant structural alteration (the average RMSd before and after the optimization is 0.03 ± 0.01 Å). In summary, we can conclude then that for most practical purposes in the field of nucleic acids simulations, original and compressed files provide the same structural information. However, the size of the compressed files is **1.4%** (95% variance cutoff) and **6.4%** (99% variance cutoff) that of the original ones.

In the preceding analysis, the attempt has been made to approximate an entire trajectory as a single set of major eigenvectors perturbing a single average structure. We hypothesized that the use of different reference coordinates and different eigenvector sets for separate sections of the whole trajectory might increase the accuracy of the procedure, especially in cases where radically different conformational states are sampled. However, for the 70 ns trajectory of DNA considered here no relevant gain (average reduction of the RMSd error of the compressed trajectory of 0.03 Å) was obtained when the trajectory was divided in 7 blocks of 10 ns and the projection→back-projection process was repeated using eigenvectors/eigenvalues and average structures were computed for each separate block. Clearly, this situation might change for systems with a less well equilibrated trajectory.

As discussed above, important eigenvectors theoretically define low-frequency movements. This opens up the possibility of using a very aggressive compression procedure where essential movements are recorded for intervals much larger than the original coordinate collection rate, and Cartesian coordinates for intermediate points are regenerated when required by linear interpolation of the projections along the set of important eigenvectors. Results in Table 2 show that the interpolation scheme introduces an additional error in the structures obtained after the projection→back-projection procedure. However, if the interpolation is used within reasonable limits this error might be acceptable for many applications (around 0.1−0.2 Å when 1 ps data is interpolated up to 10 ps samplings). Remarkably, the size of the compressed trajectory is **0.1%** (95% cutoff) that of the original Cartesian one. Further improvements might be made if different spacings were used for collecting data for low- and high-frequency modes, but the investigation of this point falls outside the scope of this article.

Table 2. RMSd (in Å) between Original (Cartesian) and Projected→Back-Projected Coordinates (Determined Using 95% Variance Cutoff) Using Different Interpolation Schemes[a]

interpolation level (spacings)	RMSd (Cartesian)	ΔRMSd
1 ps → 1 ps	0.59 ± 0.04	
2 ps → 1 ps	0.70 ± 0.05	0.11
5 ps → 1 ps	0.72 ± 0.06	0.13
10 ps → 1 ps	0.79 ± 0.08	0.20
20 ps → 1 ps	0.89 ± 0.13	0.40

[a] In all cases the lost of quality (determined as the increase in RMSd relative to that obtained with no interpolation (1ps→1ps)) is indicated. Interpolations were carried out considering a Gaussian noise defined by a standard deviation of 10% the width of the bin.

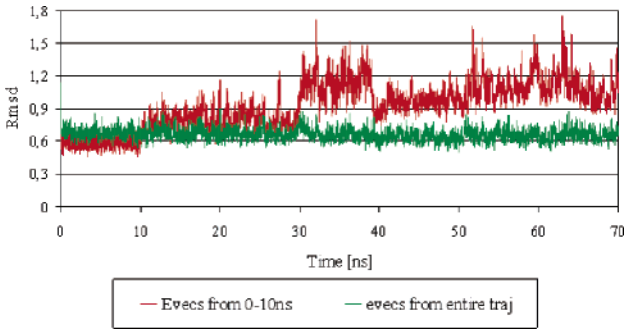


Figure 4. RMSd (in Å) between the real DNA-trajectory and coordinates generated using the projection→back-projection procedure when eigenvectors/eigenvalues and reference structures are obtained using only the first 10 ns of trajectory data. A reference profile is included (in green) indicating the expected errors when eigenvectors/eigenvalues and reference structures are obtained from analysis of the whole trajectory.

Since eigenvectors/eigenvalues describe the movements performed by a molecule along a section of its trajectory, for very long equilibrium trajectories the important movements sampled by a system in the period t-Δt→t should be identical to those sampled in the period t→$t+\Delta t$. Accordingly, in the limit of perfectly equilibrated trajectories the set of eigenvectors/eigenvalues obtained using the sampling collected in the period $[t\text{-}\Delta t,t]$ might be used to extend the trajectory to $[t,t+\Delta t]$. This is the basis of the method known as Essential Dynamics,[3−6] which is of particular interest because MD (or Monte Carlo) simulations in the essential space can be computationally very efficient. However, for the method to be reasonable it is necessary that the set of eigenvectors/eigenvalues obtained from the trajectory over the period $[t\text{-}\Delta t,t]$ can also provide an accurate representation of the essential movements for the trajectory over the period $[t,t+\Delta t]$. To investigate this point we computed the important eigenvectors for the first 10 ns of the trajectory and then used these for the projection→back-projection procedure over intervals 10−20, 10−30,...,10−70 ns. As Figure 4 reveals, the errors related to the use of eigenvectors obtained from a previous segment can be twice as large as those obtained

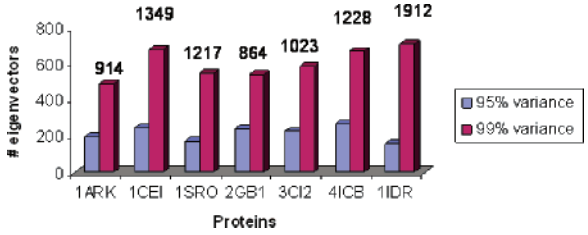


Figure 5. Number of eigenvectors needed to represent 95% or 99% of variance in the protein trajectories considered here. The number of atoms in each protein is displayed on top of the histogram bars.

when the projection→back-projection procedure is performed for the entire trajectory. The particularly mediocre performance of the approach in capturing the behavior of the system during the 30−40 ns period suggests that during this time the system underwent a form of conformational change that was not present in the 0−10 ns period and so was not captured effectively by any eigenvector used. Overall, the discrepancy between extrapolated and real trajectories increases with time, suggesting a time-dependent degradation in the quality of the eigenvectors outside their region of origin. Obviously the errors can be reduced by taking longer simulation periods for the determination of the eigenvectors. For example, if eigenvectors are obtained using the 0−30 ns simulation data, the error in the predicted coordinates over the 30−70 period reduces to 0.8−0.9 Å. In conclusion then, caution is needed in the use of essential dynamics to extend trajectories, since low-frequency movements, which are not well represented in a short-time simulation, are important to trace deformation in distant periods of time

Protein Simulations. The number of eigenvectors needed to represent the essential dynamics of proteins is larger than that of DNA duplexes of similar size. Thus, we need around 200 and between 500 and 700 essential modes to represent 95% and 99% of the variance, respectively (see Figure 5), in other words 3−5 times more eigenvectors than needed for a DNA molecule of similar size. The number of important eigenvectors does not dramatically increase when longer simulation times are used, or when the size of the protein increases (see Figure 5 obtained for 1idr). In any case, the number of important eigenvectors is still much smaller than the number of degrees of freedom of the proteins (between 2600 and 5700 for the proteins considered here), suggesting that compression should be an effective approach to reducing the size of the files.

The average all-atoms RMSd between the MD-averaged structure and the collected snapshots are between 1.2 and 2.4 Å, with the largest point deviations being above 4 Å (see as example Figure 6). When the projection→back-projection procedure is performed using only the first eigenvector (which explains between 20 and 35% of variance), the RMSd between original and back-projected conformations is reduced to 1−3 Å for the 7 proteins considered. When the space is expanded to consider the first 10 eigenvectors (around 50−60% of the total variance) the RMSd is similar or less than 1 Å for all the proteins (see as example Figure 6). The error is reduced to around 0.3 Å (10 ns trajectories) or 0.5 Å (100 ns trajectories) when the

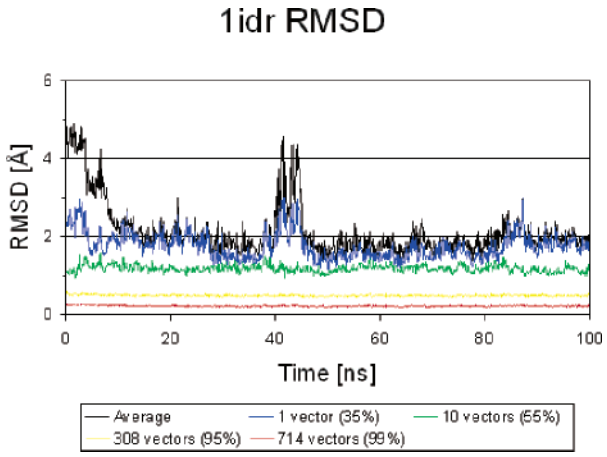


Figure 6. RMSd (in Å) between the real 100 ns trajectory of 1idr and coordinates generates using the projection→back-projection procedure with different number of eigenvectors (1: blue; 10: green; 95% variance: yellow; 99% variance: red). For reference, the results obtained when no eigenvectors are used (average structure) are also displayed (black).

Table 3. RMSd (in Å) between the Original Trajectories and Those Obtained after the Projection→Back-Projection Using a Single Reference Coordinate and the Set of Eigenvectors Necessary To Explain 95 or 99% of the Variance[a]

	95% cutoff		99% cutoff	
protein	RMSd	file size	RMSd	file size
1ark	0.36	8.5	0.15	20
1cei	0.36	7.8	0.16	20
1sr0	0.45	6.0	0.20	18
2gb1	0.29	10.0	0.13	22
3ci2	0.36	8.6	0.16	21
2icb	0.33	8.8	0.14	22
1idr	0.50	5.1	0.22	14

[a] The size of the trajectory file obtained after the projection procedure (% of the original file) is also indicated.

important eigenvectors are defined using a 95% variance cutoff and to around 0.1 (10 ns trajectories) and 0.2 (100 ns trajectory) Å when the 99% variance cutoff is used (see Table 3). As found for DNA, small geometrical errors arising from the neglect of high frequency movements can be easily corrected with a simple minimization protocol (between 20 and 50 energy minimization steps) without alteration of the global structure (RMSd below 0.03 Å).

As noted above, the compression method is exact when all the eigenvectors are considered. However, its computational efficiency should increase as trajectory behaves more harmonically. Thus, we can expect that for trajectories following irreversible transitions "nonequilibrium" trajectories the method will be less accurate. In a practical test we compare two 10 ns segments of a trajectory, the first showing a fast irreversible transition, and the second corresponding to an equilibrium trajectory (see Figure S1 in Supporting Information). Since the first trajectory is dominated by the irreversible transition, the total number of eigenvectors needed to explain a given variance threshold decreases (for example for 99% variance a reduction of 200 eigenvectors;

Table 4. RMSd (in Å) between Original (Cartesian) and Projected→Back-Projected Coordinates (Determined Using 95% Variance Cutoff) Using Different Interpolation Expansions for 1idr (Similar Relative Values Were Obtained for the Other Proteins)[a]

interpolation level	RMSd (Cartesian)	ΔRMSd
1 ps → 1 ps	0.51	
2 ps → 1 ps	0.68	0.17
5 ps → 1 ps	0.83	0.32
10 ps → 1 ps	0.93	0.42
20 ps → 1 ps	1.04	0.53

[a] In all cases the lost of quality (determined as the increase in RMSd from that obtained with no interpolation (1ps→1ps)) is indicated. Interpolations were carried out considering a Gaussian noise defined by a standard deviation of 10% the width of the bin.

see Table S3 in Supporting Information), and the RMSd between the real and compressed files slightly increases (for example for 99% variance from 0.15 to 0.19 Å). However, when the same number of eigenvectors is considered, the performance of the method is almost identical for the two trajectories (see Table S3 in Supporting Information).

In summary, the compression procedure provides a set of coordinates that is nearly indistinguishable (for most purposes) from the original ones. Very interestingly, the size of the compressed files is on average (see Table 3) 8% (95% variance cutoff) and 20% (99% variance cutoff) that of the original trajectories. The reduction becomes more evident for longer trajectories. Additional savings of disk space can be obtained by adding the interpolation procedure outlined above; however, it introduces an additional error which can be too large when it is performed between snapshots too far apart in time. Our results suggest (see Table 4) that a 5→1 ps expansion seems a good compromise between the reduction in the size of the files and the loss of quality in the generated coordinates. Note that this interpolation procedure reduces to 1/5 the size of the projection data, which is the only part of the compressed format which depends on the length of the trajectory.

The use of multiple reference conformations and associated eigenvectors/ eigenvalues is not justified for short (10 ns) trajectories and leads to only a modest increase in the performance of the method for long trajectories. In fact, using 10 sets (10 ns each) of references structures and eigenvectors/ eigenvalues, the RMSd error between original and compressed conformations for the 100 ns trajectories was reduced by 0.1 Å (for both 95 and 99% variance limits). We expect that the multiple-reference strategy may be more effective for more complex trajectories showing large variations in the average structure.

Finally, the use of eigenvectors obtained in a short trajectory fragment to describe the movements in more distant regions of the trajectory leads to non-negligible errors in the back-projected coordinates with respect to the real ones and also to those generated by the usual compression procedure (see Figure 7). It is then clear that both the use of extrapolation techniques based on the sampling of essential movements defined in short simulation times must be done with caution.

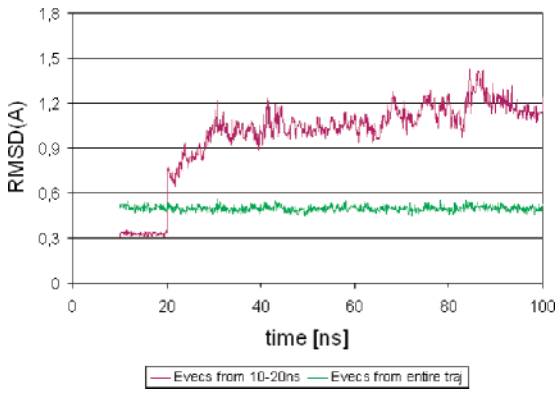


Figure 7. RMSd (in Å) between the real 1idr-trajectory and coordinates generated using the projection→back-projection procedure when eigenvectors/eigenvalues and reference structures are obtained using only the first 10 ns of trajectory data. A reference profile is included (in green) indicating the expected errors when eigenvectors/eigenvalues and reference structures are obtained from analysis of the whole trajectory.

Conclusions

We find that a data compression method based on principal component analysis can work remarkably well with MD trajectory data, permitting files to be reduced to typically less than one tenth of their original size with very acceptable levels of approximation. We would suggest a file format based on this approach configured as follows (Figure 1). The file would begin with the coordinates of the time-averaged structure from the trajectory (3*N* floating point numbers). Next would come the first eigenvector (again 3*N* floats). Next would come the *T* coefficients of this eigenvector over the trajectory. The format then repeats: the second eigenvector then the second time series of coefficients, then the third, etc. The advantage of this approach is that, if one imagines the data being transmitted from one place to another, transmission may be interrupted at any point according to the accuracy required for the regenerated Cartesian coordinates.

Further work remains to be done concerning the possibility for further data compression by allowing interpolation. While in theory the major eigenvectors should relate to low frequency modes, which should be able to be accurately recreated by interpolation between sparse samplings, in practice this is not really the case. A good example of this can be seen in our recent work[19] contrasting the dynamical behavior of a DNA duplex in simulations undertaken with an implicit solvation model compared with those undertaken (as here) with explicit solvent. As Figure 4 in ref 19 shows, the effect of solvent is to contaminate the low-frequency modes with high frequency 'noise' from solvent−solute collisions. In future work we will address the question of whether it is possible to optimize interpolation schemes by a careful analysis of this phenomenon,

Acknowledgment. This work was supported by the Instituto Nacional de Bioinformática (INB-Genoma España), the Spanish Ministry of Education and Science (BIO2003-06848 and SAF2002-04282), and the Barcelona Supercomputing Center.

Supporting Information Available: Analysis of the performance of the compression procedure to reproduce helical parameters stacking and hydrogen bonding in DNA simulations and of the quality of the method to reproduce "nonequilibrium" MD simulations of proteins. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Bishop, T. C. *J. Biomol. Struct. Dyn.* **2005**, *22*, 673−686.

(2) Sanbonmatsu, K. Y.; Joseph, S.; Tung, C. S. *Proc. Natl. Acad. Sci*. **2005**, *102*, 15854−9

(3) Amadei, A.; Linsen, A. B. M.; Berendsen, H. J. C. *Proteins* **1993**, *17*, 412.

(4) Groot, B. I. de; Hayward, S.; van Aalten, D. M. F.; Amadei, A.; Berendsen, H. J. C. *Proteins* **1998**, *31*, 116.

(5) Wlodek, S. T.; Clark, T. W.; Scott, L. R.; McCammon, J. A. *J. Am. Chem. Soc.* **1997**, *119*, 9513.

(6) Orozco, M.; Pérez, A.; Noy, A.; Luque, F. J. *Chem. Soc. Rev.* **2003**, *32*, 350−364.

(7) Noy, A. Meyer, T.; Rueda, M.; Ferrer, C.; Valencia, A.; Perez, A.; de la Cruz, X.; López, J. M.; Luque, F. J.; Orozco, M. *J. Biomol. Struct. Dyn.* **2005**, in press .

(8) AMBER8.0 Computer Program. Case, D. A. et al. University of California, San Francisco, 1999.

(9) Anderson, E.; Bai, Z.; Bischof, C.; Demmel, J.; Dongarra, J.; Du Croz, J.; Greenbaum, A.; Hammarling, S.; McKenney, A.; Ostrouchov, S.; Sorensen, D. *LAPACK Users' Guide*, 2nd ed.; SIAM: Philadelphia, PA, 1995.

(10) Lehoucq, R. B.; Sorensen, D. C.; Yang, C. *ARPACK Users' Guide: Solution of Large Scale Eigenvalue Problems with Implicitly Restarted Arnoldi Methods*; SIAM: Philadelphia, PA, 1997.

(11) Morreale, A.; de la Cruz, X.; Meyer, T.; Gelpí, J. L.; Luque, F. J.; Orozco, M. *Proteins* **2004**, *58*, 101−109.

(12) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. *J. Am. Chem. Soc.* **1995**, *117*, 11946−11975.

(13) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926−935.

(14) Darden, T. A.; York, D. M.; Pedersen, L. G. *J. Chem. Phys*. **1993**, *98*, 10089−10092.

(15) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *23*, 327−341.

(16) Phillips, J.C.; Braun, R.; Wang, W.; Cumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R.D.; Kale, L.; Schulten, K. *J. Comput.Chem.* **2005**, *26*, 1781−1802.

(17) Ciccotti, J. P. G.; Berendsen, H. J. C. *J. Comput. Phys*. **1977**, *23*, 327−341.

(18) Pérez, A.; Blas, J. R.; Rueda, R.; López-Bes, J. M.; de la Cruz, X.; Orozco, M. *J. Chem. Theory Comput.* **2005**, *1*, 790−800.

(19) Sands, Z. A.; Laughton, C. A. *J. Phys. Chem. B* **2004**, *108*, 10113−10119.

CT050285B

JCTC Journal of Chemical Theory and Computation

Sensitivity Analysis and Charge-Optimization for Flexible Ligands: Applicability to Lead Optimization

Michael K. Gilson*

Center for Advanced Research in Biotechnology, University of Maryland Biotechnology Institute, 9600 Gudelsky Drive, Rockville, Maryland 20850

Received September 9, 2005

Abstract: Sensitivity analysis and charge-optimization have been suggested as methods to guide the optimization of lead compounds in early-stage drug discovery. However, applications to date have been restricted by the simplifying assumption of a rigid ligand. The present study applies both formalisms to the case of a flexible ligand in a model application to an HIV-protease inhibitor. The results suggest that sensitivity analysis is a fast and robust method for guiding charge changes in both a rigid and a flexible ligand, although its accuracy is limited by the fact that it represents a linear approximation. The more complete quadratic analysis provided by charge-optimization produces unexpected results when the ligand is considered to be flexible. For example, it can yield atomic charges which powerfully stabilize the bound conformation of the ligand relative to the conformation assumed for the free state, thus markedly destabilizing the assumed free conformation. Such results are traceable to the fact that the energy matrix possesses negative eigenvalues. However, optimizing charges under the assumption that the ligand does not change conformation upon binding leads to a set of charges that robustly improve affinity, even when the free conformation is later allowed to vary. Thus, both sensitivity analysis and charge-optimization appear to be useful techniques.

1. Introduction

Structure-based drug discovery frequently begins with identification of a lead compound, a small molecule with moderate affinity for a targeted protein of known structure. This is followed by modification of the lead compound in order to arrive at a high-affinity drug candidate. This second step, improving on the lead compound, remains a significant challenge because it is rarely clear what chemical changes will lead to greater affinity for the target. One rather obvious approach is to make changes that will increase the favorable Coulombic interactions between the ligand and the protein. However, any charges that are added to the ligand and that come to lie in the ligand−protein interface are stripped of water during the process of binding and thus incur a desolvation free energy penalty which can be substantial. In fact, calculations frequently indicate that the desolvation penalty exceeds the attractive interaction, making for an unfavorable net electrostatic interaction even when the binding interface appears to possess good electrostatic complementarity; see, e.g. refs 1 and 2.

In recent years, a series of studies (see, e.g., refs 3−5) has addressed this issue, pointing out that the variation of the electrostatic contribution to the binding energy as a function of the atomic charges of the ligand has the form of a parabola with upward curvature, when the ligand is considered to be rigid. As a consequence, there is a set of ligand charges that minimizes the change in electrostatic energy and hence maximizes affinity, all other things being equal. Moreover, when optimal or near-optimal charges are assumed, the net change in electrostatic energy upon binding can be strongly favorable. These observations have led to the exploration[5−7] and successful use[6,8] of charge-optimization methods as a basis for lead optimization.

One potential limitation of the charge optimization approach is that the values of the optimal charges in one part of a ligand are influenced by charges assumed to exist at

* Corresponding author phone: (240)314-6217; fax: (240)314-6255; e-mail: gilson@umbi.umd.edu.

10.1021/ct050226y CCC: $33.50
Published on Web 01/27/2006

other ligand atoms. As a consequence, if one is interested in identifying only parts of the ligand whose charges can be changed to improve affinity, the charges of a fully optimized ligand may not faithfully indicate the changes needed for just a part, as recently noted[9] and further discussed in this paper. Another potential drawback of the method is that it tends to be time-consuming, at least as originally formulated, since the method has required solving the linearized Poisson−Boltzmann (LPB) equation at least once for every atom whose charge is to be optimized. This means on the order of 100 LPB calculations for a druglike ligand. On the other hand, recent algorithmic advances promise to markedly reduce the computational cost of the method.[10,11]

Sensitivity analysis represents another promising approach to guiding the electrostatic optimization of ligands; see, e.g. refs 9 and 12−14. In the present context, this method involves computing the first derivative of the binding free energy with respect to the partial atomic charges of the ligands; affinity can then be improved by raising the charge of atoms with negative derivatives or lowering the charge of atoms with positive derivatives. Sensitivity analysis requires fewer numerical solutions of the LPB equation than does standard charge-optimization and therefore should be less time-consuming. On the other hand, it is a linear approximation and thus does not account for the parabolic curvature of the electrostatic energy with respect to atomic charges. As a consequence, it may be less accurate.

Thus, both charge-optimization and sensitivity analysis are theoretically interesting methods that could be quite useful during the lead optimization stage of drug discovery. To date, however, these methods have been limited to applications in which the ligand is assumed to be rigid. In fact, it has not been clear whether charge-optimization could be generalized to the more realistic case of a flexible ligand. Moreover, we are not aware of a direct comparison of the two approaches, for either a rigid or a flexible ligand. The present paper thus discusses how both methods generalize to the case of a flexible ligand and compares their properties for a model system in which the ligand is treated first as rigid and then as flexible.

2. Theory

This section derives the equations of sensitivity analysis and charge-optimization from statistical thermodynamic expressions for the binding affinity of a ligand and a protein, allowing for molecular flexibility. The resulting expressions prove to be essentially the same as for a rigid ligand, except that Boltzmann-averaged electrostatic potentials replace the potentials computed for a single conformation of the ligand. The rigid ligand thus represents a special case in which the Boltzmann average includes only a single conformation having a probability of 1. However, accounting for ligand flexibility can strongly affect the results obtained with these methods, especially charge-optimization.

2.1. Derivatives of the Binding Free Energy. The standard free energy of association of a ligand L and receptor R to form a noncovalent complex can be written as the difference between the standard chemical potentials μ° of the respective molecular species

$$\Delta G^\circ = \mu^\circ_{RL} - \mu^\circ_R - \mu^\circ_L \quad (1)$$

where R, L, and RL indicate the receptor, the ligand, and their complex, respectively. The standard chemical potential of each molecular species in turn can be written as[15,16]

$$\mu^\circ_X = -RT \ln\left(\frac{8\pi^2}{C^\circ}\int e^{-\beta E(\mathbf{s},\mathbf{r})} d\mathbf{r}\right) \quad (2)$$

Here C° is the standard concentration which, combined with the factor of $8\pi^2$, accounts for the positional and orientational mobility of the free molecule at standard concentration;[15] $\beta \equiv 1/kT$, k being Boltzmann's constant and T the absolute temperature; $\mathbf{r}$ represents the internal coordinates of the molecular species and thus defines its three-dimensional conformation; and $E(\mathbf{s}, \mathbf{r})$ is the energy of the molecule as a function of its conformation and a set of computational parameters $\mathbf{s}$. In a typical force field-based calculation, $\mathbf{s}$ will include such solute parameters as atomic partial charges and Lennard-Jones parameters, along with solvent parameters such as the atomic partial charges of an explicit water model like TIP3P or the dielectric constant of an implicit solvent model. The derivative of the chemical potential with respect to atomic parameter s_i is

$$\frac{\partial \mu^\circ_X}{\partial s_i} = \frac{\int \frac{\partial E(\mathbf{s},\mathbf{r})}{\partial s_i} e^{-\beta E(\mathbf{s},\mathbf{r})} d\mathbf{r}}{\int e^{-\beta E(\mathbf{s},\mathbf{r})} d\mathbf{r}} = \left\langle \frac{\partial E(\mathbf{s},\mathbf{r})}{\partial s_i} \right\rangle \quad (3)$$

The quantity in angle brackets is the Boltzmann average of the derivative of the energy function with respect to the parameter of interest. This formula is consistent with prior expressions for free energy derivatives from Cieplak and coworkers.[17]

In the present application, we are interested in the case where the energy model includes a solvent-screened charge−charge interaction term and a solvent reaction field term, both linear with respect to the N atomic charges $\mathbf{q} \equiv (q_1, q_2, \ldots, q_N)$, along with other energy contributions that do not depend directly upon atomic charges. The other contributions can be lumped together as $E_{other}(\mathbf{s},\mathbf{r})$, where $\mathbf{s}$ now refers exclusively to noncharge parameters. The energy thus can be written as

$$E(\mathbf{q},\mathbf{s},\mathbf{r}) = \sum_{i=1}^{N}\sum_{j>i}^{N} \frac{q_i q_j}{D^{eff}_{ij} r_{ij}} + \frac{1}{2}\sum_{i=1}^{N} q_i \phi^{RF}_i + E_{other}(\mathbf{s},\mathbf{r}) \quad (4)$$

where i and j index the molecule's atoms, D^{eff}_{ij} is the effective dielectric constant[18] of the interaction between atoms i and j, and ϕ^{RF}_i is the part of the solvent reaction field at atom i that is induced by the charge at atom i. The assumption of linearity implies furthermore that

$$\phi^{RF}_i \propto q_i \quad (5)$$

Note that the proportionality constant and the values of D^{eff}_{ij}

and r_{ij} depend on the atomic coordinates $\boldsymbol{r}$; i.e., upon the conformation of the molecule.

Substituting eq 5 into eq 4 and then using eq 3 to take the derivative of the chemical potential with respect to q_i yields that

$$\frac{\partial \mu^\circ}{\partial q_i} = \left\langle \sum_{j \neq i}^{N} \frac{q_j}{D_{ij}^{\mathrm{eff}} r_{ij}} + \phi_i^{\mathrm{RF}} \right\rangle = \langle \phi_i \rangle \tag{6}$$

Thus, the derivative of the molecule's chemical potential with respect to the charge of atom i is the Boltzmann-averaged electrostatic potential at atom i. This observation is not surprising, but it is useful because it generalizes what is commonly known for a rigid molecule to the case of one that is flexible. The rigid molecule becomes a special case of eq 6, in which the Boltzmann average of the potential equals the potential computed for a single conformation.

Finally, eq 6 can be combined with eq 1 to show that the derivative of the binding free energy with respect to the charge of atom i, which can belong to either the ligand or the receptor, is simply

$$\frac{\partial G^\circ}{\partial q_i} = \langle \phi_i^{\mathrm{b}} \rangle - \langle \phi_i^{\mathrm{f}} \rangle \tag{7}$$

where the superscripts "b" and "f" indicate respectively the bound (RL) and free (R or L) states of the system. Sensitivity analysis can then be used to generate a first-order prediction of the change in binding energy for a small change in an atomic charge, Δq_i:

$$\Delta G \approx \frac{\partial \Delta G}{\partial q_i} \Delta q_i \tag{8}$$

The present analysis is consistent with an earlier and more general discussion of the application of sensitivity analysis to free energies.[12]

2.2. Electrostatic Optimization. The theory of electrostatic optimization for rigid molecules has been elegantly laid out and explored by its originators; see, e.g., refs 3 and 4. Here, the binding energy derivatives discussed in section 2.1 are employed to generalize the formalism of electrostatic optimization to the case of flexible molecules.

Consider a ligand L with N atoms of charge q_i, $i \in [1...N]$ and a receptor R with M atoms of charge q_i, $i \in [N+1...N+M]$. We wish to find the values of the ligand charges that minimize the binding energy and thus maximize affinity, subject to the physically reasonable constraint that the total charge of the ligand, Q, equals a user-specified integer; i.e., that

$$g(\boldsymbol{q}) = \sum_{i}^{N} q_i - Q = 0 \tag{9}$$

Equation 9 introduces the function $g(\mathbf{q})$ which is used to define the constraint on total charge. A set of charges that minimizes the binding energy subject to this constraint can be found by the method of Lagrangian multipliers (e.g., ref 19 p 946), which leads to a system of N equations, one for each atomic charge q_i

$$\frac{\partial \Delta G^\circ}{\partial q_i} + \lambda \frac{\partial g}{\partial q_i} = \frac{\partial \Delta G^\circ}{\partial q_i} + \lambda = 0 \tag{10}$$

where λ, the undetermined multiplier, represents an additional unknown. Supplementing eqs 10 with the constraint on the total charge, eq 9, yields a system of $N+1$ equations in $N+1$ unknowns. It is important to note that, although eq 10 is a necessary condition for a set of charges to minimize the binding free energy while meeting the constraint on total charge, it is not sufficient to guarantee an energy-minimum, because a set of charges that satisfies eq 10 could also be a maximum or a saddle, at least in principle. The nature of the stationary point will be determined by the specific molecular problem and the parameters of the calculation.

Interestingly, eq 10 can be rewritten with the aid of eq 7 as

$$\langle \phi_i^{\mathrm{b}} \rangle - \langle \phi_i^{\mathrm{f}} \rangle = -\lambda \tag{11}$$

This implies that the charges from the Lagrangian procedure also cause all atoms of the ligand to experience the same change in mean potential upon binding and that this change in potential is $-\lambda$. This equation generalizes the concept of a residual potential at each atom[4] which goes to zero for a ligand which minimizes the energy in the absence of any constraint on the total charge. Equation 11 shows that the residual potential goes to $-\lambda$ rather than zero when the constraint is imposed.

Assuming the classical properties of linearity and reciprocity allows eqs 10 to be rewritten as a set of N linear equations. Working from eqs 5 and 7, we define

$$a_{ii} \equiv \frac{\langle \phi_i^{\mathrm{RF}} \rangle}{q_i} \tag{12}$$

$$a_{ij} \equiv \langle (D_{ij}^{\mathrm{eff}} r_{ij})^{-1} \rangle \tag{13}$$

where a_{ii}^{b}, a_{ij}^{b} and a_{ii}^{f}, a_{ij}^{f} refer to values for the bound and free states of the ligand, respectively. Substituting these terms into eq 6 and then using eq 11 yields

$$q_i(a_{ii}^{\mathrm{b}} - a_{ii}^{\mathrm{f}}) + \sum_{j \neq i}^{N} q_j(a_{ij}^{\mathrm{b}} - a_{ij}^{\mathrm{f}}) + \sum_{j=N+1}^{N+M} q_j(a_{ij}^{\mathrm{b}} - a_{ij}^{\mathrm{f}}) + \lambda = 0 \tag{14}$$

for $i = (1, 2, \ldots, N)$. The first term is the change upon binding of the reaction field at atom i due to charge q_i; the second term is the change upon binding of the screened Coulomb potential at atom i due to other ligand atoms j; and the third term is the change in the potential at atom i of the ligand due to the charges j of the receptor, ϕ_i^{b}. Note that the ligand does not feel the receptor when they are not bound, so $a_{ij}^{\mathrm{f}} = 0$.

Equations 9 and 14 can be expressed together in matrix form

$$\begin{pmatrix} a_{11}^{b} - a_{11}^{f} & a_{12}^{b} - a_{12}^{f} & \cdots & a_{1N}^{b} - a_{1N}^{f} & 1 \\ a_{21}^{b} - a_{21}^{f} & a_{22}^{b} - a_{22}^{f} & \cdots & a_{2N}^{b} - a_{2N}^{f} & 1 \\ & & \cdots & & \\ a_{N1}^{b} - a_{N1}^{f} & a_{N2}^{b} - a_{N2}^{f} & \cdots & a_{NN}^{b} - a_{NN}^{f} & 1 \\ 1 & 1 & \cdots & 1 & 0 \end{pmatrix} \begin{pmatrix} q_1 \\ q_2 \\ \cdots \\ q_N \\ \lambda \end{pmatrix} = \begin{pmatrix} -\phi_1^{b} \\ -\phi_2^{b} \\ \cdots \\ -\phi_N^{b} \\ Q \end{pmatrix} \tag{15}$$

or

$$\mathbf{Aq} = \mathbf{B} \tag{16}$$

The equation represented in the lowest row of the matrices is the constraint on the total charge (eq 9). A typical druglike ligand possesses on the order of 100 atoms, so the dimensions of $\boldsymbol{A}$ are roughly 100 × 100. This matrix equation can be readily solved by matrix inversion or by iterative methods to yield a stationary point $\mathbf{q}^\circ$ on the hyperplane of net charge Q, along with the corresponding value of λ which, as noted earlier in this section, equals the change in potential on binding when the charges equal $\mathbf{q}^\circ$.

The change in the solvation and the intramolecular charge−charge interactions of the ligand upon binding is $\mathbf{q}^{\circ T}\mathbf{A}\mathbf{q}^\circ$, while the interaction of the ligand with the protein is $\mathbf{q}^{\circ T}\mathbf{B}$, where $\mathbf{q}^\circ \equiv (q_1, q_2, \ldots, q_N, 0)$. The change in the protein solvation energy upon binding also contributes to the overall electrostatic binding energy, but this quantity is independent of the ligand charges. Therefore, charge optimization focuses on the change in ligand−ligand and ligand−protein energies:

$$\Delta G_{ll,lp} = \mathbf{q}^{\circ T}(\mathbf{A}\mathbf{q}^\circ + \mathbf{B}) \tag{17}$$

The full change in electrostatic energy can be obtained by separately computing the change in protein solvation energy, ΔG_{pp}, and adding it to $\Delta G_{ll,lp}$.

3. Methods

3.1. Molecular Systems and Parameters. Charge optimization and sensitivity analysis were studied for a model system, the association of HIV-1 protease with the cyclic urea inhibitor XK263.[20] The protein was fixed in its crystal conformation,[21] and a single conformation of the bound ligand was considered. However, three conformations were considered for the free ligand: a conformation identical to the bound conformation and two alternate conformations. Thus, the binding processes considered are as follows:

Rigid Ligand: $\text{LigConf}^0 + \text{Protein} \rightarrow \text{LigConf}^0\cdot\text{Protein}$

Flexible Ligand 1: $\text{LigConf}^1 + \text{Protein} \rightarrow \text{LigConf}^0\cdot\text{Protein}$

Flexible Ligand 2: $\text{LigConf}^2 + \text{Protein} \rightarrow \text{LigConf}^0\cdot\text{Protein}$

Here LigConf^0 is the bound conformation of the ligand, and LigConf^1 and LigConf^2 are the two alternate conformations of the free ligand, which will be referred to as Flex 1 and Flex 2.

The molecular models were prepared as follows. Polar hydrogen atoms were added to the crystal structure, and CHARMM[22] force-field parameters were assigned with the program Quanta.[23] The cyclic urea inhibitor was relaxed by redocking it to the original crystal structure of the protease with the program Vdock,[24,25] which allows continuous variation of nonring single-bonds and of the overall position and orientation of the compound within the binding site. The resulting conformation (LigConf^0) was used in calculating electrostatic terms for the receptor−ligand system and for the free ligand when it was considered as rigid. The two alternative conformations of the free ligand, Flex 1 and Flex 2, were generated by running on the order of 10−100 ps of stochastic dynamics[26] at 300 K for the ligand alone, with a time-step of 1 fs. During the MD calculations, screening of electrostatic interactions by solvent was accounted for in an approximate fashion via the distance-diependent dielectric model with a coefficient of 4.

Electrostatic terms were obtained from finite-difference solutions of the linearized Poisson−Boltzmann equation carried out with the program UHBD,[27] as detailed in section 3.2. The dielectric constants of the solutes and solvent were set to 1 and 78.5, respectively, with solvent ionic strength of 150 mM and an ion-exclusion radius (Stern layer) of 2 Å thickness. The boundary between the low dielectric interior and the high dielectric solvent was defined by the molecular surface[28] with a probe radius of 1.4 Å and atomic radii set to their CHARMM Lennard-Jones R_{min} values.

3.2. Calculation of the Electrostatic Terms. To set up the electrostatic optimization problem for a given system, it is necessary to obtain the values of a_{ij}^{b}, a_{ij}^{f}, and ϕ_i^{b}, where i, $j \le N$. (See eq 14.) These terms were obtained here with a series of finite-difference Poisson−Boltzmann[29−31] (FDPB) calculations. In each case, potentials were computed via an initial FDPB calculation with a coarse grid spacing of 0.5 Å and then a second "focusing"[31] FDPB calculation with a grid spacing of 0.2 Å, where the grid encompassed the entire ligand, and boundary conditions were drawn from the initial coarse grid run. These quantities were computed either under the assumption of a rigid ligand or a flexible ligand whose conformation changed from one bound conformation to a single different free conformation. However, as discussed in the Theory section, it would also be possible to compute Boltzmann averages over multiple bound and/or free conformations.

The **B** vector was filled with values of ϕ_i^{b} by a single FDPB calculation for the ligand−receptor complex in which all ligand charges were artificially set to zero, but all protein charges were kept at their normal values. The resulting electrostatic potential at atom i is ϕ_i^{b}.

The **A** matrix was filled by computing the values of a_{ij}^{b} and a_{ij}^{f}; see eqs 12 and 13. The calculations for a_{ij}^{b} are the same as those for a_{ij}^{f} except that the former use the bound conformation of the ligand in the presence of the protein, which is treated as electrically neutral, while the latter use the free conformation of the ligand in the absence of the protein. (As noted above, the free conformation is the same as the bound conformation when the ligand is assumed to be rigid.) Hence, the following description refers to a_{ij}^{b} and

a_{ij}^{f} generically as a_{ij}. The values of a_{ij} were computed as the sum of Coulombic potentials ϕ_{ij}^{C} based on the dielectric constant of the protein interior and reaction field potentials produced by the solvent. The Coulombic potentials were computed by placing a unit charge on atom i and zeroing all other charges, setting the dielectric constant of the solvent region to the dielectric constant of the molecular interior, and evaluating the resulting potentials at every other atom. Self-interactions ($i = j$) were omitted, along with interactions between atoms directly bonded to each other and interactions between atoms in a 1−3 bonding relationship. These exclusions are standard practice in force field calculations, and here they avoid allowing nonphysical short-ranged interactions to influence the charge optimization procedures. Interactions across a dihedral angle (1−4 interactions) were included without any special scaling factor. The solvation parts of a_{ij} were computed by carrying out an additional FDPB calculation with a unit charge on atom i and all other charges zeroed but now with the solvent dielectric constant set to the solvent value. The solvation parts of a_{ij} were then set to the difference between the solvated and the unsolvated potential at atom j. Note that this same procedure gives the correct value for a_{ii}.

3.3. Numerical Methods. Matrices were diagonalized with dsyev and associated subroutines, and matrix equations were solved with dgesv and associated subroutines, all drawn from LAPACK.[32] In prior applications of Lagrangian charge optimization, the values of q_i have been constrained to lie within a range that is typical for current empirical force fields; e.g., $q_i \leq 0.85$. No such constraint was applied here, however.

When the top left $N \times N$ submatrix of the **A** matrix possesses negative eigenvalues, the charges provided by the method of Lagrangian multipliers may not correspond to a minimum of the electrostatic energy; they could also represent a saddle point or a local maximum. In such cases, charges that minimize the electrostatic energy were sought with the minimization program PRAXIS,[33] obtained from the Netlib repository of mathematical software.[34] More particularly, PRAXIS was used to minimize a quantity consisting of $\Delta G_{ll,lp}$ (eq 17) supplemented with a pseudoenergy term which restrains the absolute value of the total charge to zero. In some calculations, additional pseudoenergy terms were included to keep individual atomic charges in the range −0.85 to 0.85, as previously proposed.[6] Note that $\Delta G_{ll,lp}$ can be computed very quickly for a given set of ligand charges by using the precalculated **A** and **B** matrices in eq 17, so these minimizations are not overly time-consuming. The results of minimization with the PRAXIS algorithm can depend on the initial guess for the values of the charges.

4. Results

This section describes the properties of sensitivity analysis and charge-optimization when the free conformation of the ligand is assumed to be the same as that of the bound conformation and when the free ligand is considered to adopt a different conformation.

4.1. Rigid Ligand. 4.1.1. Sensitivity Analysis. Table 1, columns 2 and 4, shows the derivatives $\partial\Delta G/\partial q_i$ of the binding energy with respect to each atomic charge i when the free ligand conformation is considered to be the same as the bound conformation; i.e., for the assumption of a rigid ligand. Atom labels are listed in Figure 1. Nearly all the derivatives are negative, indicating that making the atomic charge more positive will favor binding. This is a conse-

Table 1. Derivatives of Binding Free Energy with Respect to Partial Atomic Charges of Ligand Atoms for Rigid Ligand (kcal/mol/au), i.e., with Free Conformation of Ligand Same as Bound Conformation

atom i	$\partial\Delta G°.\partial q_i$	atom i	$\partial\Delta G°.\partial q_i$	atom i	$\partial\Delta G°.\partial q_i$
C1	−26.58	C37	−42.51	H11	−37.28
O1	−7.63	C61	−59.54	H12	−28.18
N2	−38.44	C62	−39.58	H13	−16.18
C2	−23.81	C63	−36.87	H14	−8.10
C3	−67.09	C64	−11.40	H15	−7.12
C4	−102.04	C65	−10.09	H16	−12.06
O4	−129.84	C66	−12.74	H17	−27.51
C5	−97.85	C67	−21.12	H18	−47.88
O5	−120.70	C70	−31.19	H19	−88.61
C6	−63.66	C71	−29.86	H20	−10.71
N7	−37.24	C72	−26.47	H21	2.03
C7	−23.48	C73	−18.29	H22	−3.00
C20	−32.88	C74	−14.76	H23	−7.53
C21	−36.87	C75	−15.77	H24	−60.23
C22	−32.66	C76	−18.91	H25	−47.69
C23	−27.48	C77	−26.61	H26	−85.14
C24	−17.72	C78	−29.91	H27	−55.48
C25	−11.92	C79	−32.81	H28	10.81
C26	−13.10	H1	−11.06	H29	−2.74
C27	−18.08	H2	−15.40	H30	−3.06
C28	−27.12	H3	−67.54	H31	−12.00
C29	−30.08	H4	−103.80	H32	−29.57
C31	−61.58	H5	−150.63	H33	−12.10
C32	−41.25	H6	−94.64	H34	−8.23
C33	−19.72	H7	−142.94	H35	−19.07
C34	−10.02	H8	−62.18	H36	−16.80
C35	−11.24	H9	−11.18	H37	−25.66
C36	−19.85	H10	−16.86	H38	−31.35

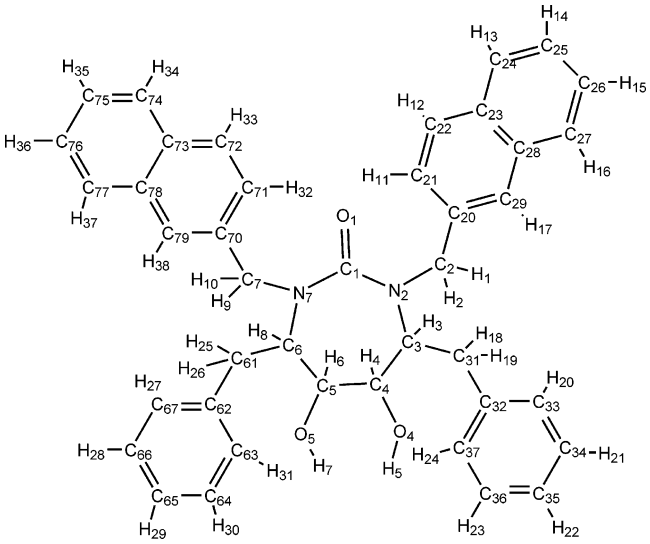


Figure 1. Diagram of ligand XK263 with atom codes used in tables.

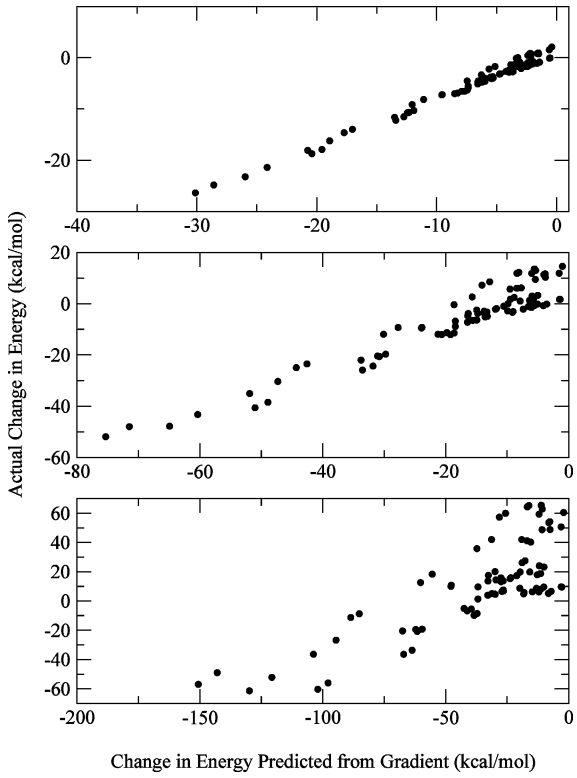


Figure 2. Accuracy of energy predictions from sensitivity analysis when the ligand is assumed rigid, shown as scatter plots of the change in electrostatic energy (kcal/mol) computed with the full parabolic energy surface versus the change predicted by sensitivity analysis, for charge changes of 0.2 au (top), 0.5 au (middle), and 1.0 au (bottom).

quence of the dominant influence of the two negatively charged aspartyl groups in the active site of the protease, which produce a positive potential at virtually every atom of the ligand.

The accuracy of sensitivity analysis is assessed here by changing each ligand charge by a small amount Δq_i in a direction opposite to the local derivative and using eq 8 to estimate the resulting change in binding energy. Figure 2 compares these first-order predictions with the actual value of $\Delta\Delta G_{ll,lp}$ computed with eq 17. Comparisons are shown for charge changes of 0.2 au (top), 0.5 au (middle), and 1.0 au (bottom). It is evident that the smaller charge changes almost always improve the binding energy, and the linear predictions from the gradients correlate well with the actual results. However, for charges changes of 1.0 au, many of the charge changes make the binding energy more positive, and the energy predictions are quite poor. This is a consequence of the quadratic dependence of energy upon charge: large charge changes often cross the energy minimum and climb the far side of the parabola.

4.1.2. Charge Optimization. The eigenvalues associated with the top left $N \times N$ submatrix of **A** obtained when the ligand is assumed to be rigid are plotted in ascending order in Figure 3 (black line). All eigenvalues are positive, indicating that the charges from the method of Lagrangian multipliers will represent not only a stationary point of the binding energy but also a minimum. (A maximum is mathematically possible but unlikely physically because the

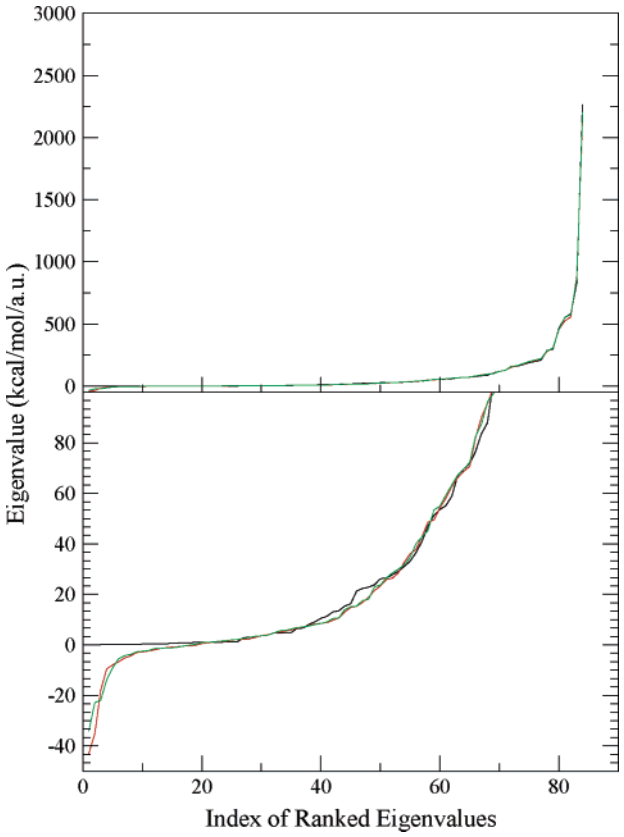


Figure 3. Rank-ordered eigenvalues of top left $N \times N$ submatrix of **A** matrix when ligand is assumed rigid (black), and when two variant conformations of the free ligand are assumed (red, green). (See Flex 1 and Flex 2 data in Table 2.) The bottom graph is the same as the top graph except for the scale of the ordinate.

desolvation energy of the ligand upon binding, which depends quadratically upon charge, is expected to be unfavorable, leading to a parabolic energy function with upward curvature.) Table 2 (columns 3 and 10) lists these optimal charges. Although the individual charges are not constrained, only a few are larger than 1 au in magnitude and none are greater than 1.6 au. Table 2 (columns 2 and 9) lists the ligand charges assigned by Quanta/CHARMM, for comparison. Not surprisingly, there is no evident correlation with the optimal charges in columns 2 and 9.

Table 2 also compares $\Delta G_{ll,lp}$, the change in electrostatic energy upon binding less the protein desolvation energy, for the optimized charges and the CHARMM charges. Optimizing the charges dramatically lowers the energy, from −15.1 kcal/mol to −113 kcal/mol. The full change in electrostatic energy can be obtained by adding the cost of desolvating the protein, ΔG_{pp}; a separate calculation yields a value of 98.4 kcal/mol for this final part of the energy, yielding a net change in electrostatic energy of −14.6 kcal/mol. These results are consistent with previous work noting that, although continuum electrostatics models tend to yield an unfavorable binding energies, the electrostatic energy can contribute favorably to binding if the charges are right.[3,4]

Like sensitivity analysis, the charge optimization methodology can be used to guide chemical modifications of a ligand aimed at improving its affinity. Figure 4 evaluates

Table 2. Partial Atomic Charges (au, in Top 42 lines) and Associated Electrostatic Contributions to the Binding Free Energy (kcal/mol) for Three Assumptions about the Conformation of the Ligand in the Free State (Bottom 3 Lines on Right-Hand Side)[a]

atom	CHARMM	optimized (rigid)	stationary point (Flex 1)	stationary point (Flex 2)	min. 1 (Flex 1)	min. 2 (Flex 1)	atom	CHARMM	optimized (rigid)	stationary point (Flex 1)	stationary point (Flex 2)	min. 1 (Flex 1)	min. 2 (Flex 1)
C1	0.600	1.559	−0.008	−1.647	−0.850	0.850	C78	−0.129	−0.121	8.005	−2.198	0.336	0.360
O1	−0.550	−0.723	−1.086	0.154	−0.846	−0.061	C79	−0.129	0.173	−1.478	2.410	−0.846	0.148
N2	−0.250	−1.699	3.928	2.835	−0.845	0.239	H1	0.051	−0.004	−0.181	−0.051	0.791	−0.850
C2	0.091	0.566	−3.503	−0.640	−0.281	0.479	H2	0.051	−0.187	−0.070	0.148	0.850	−0.850
C3	−0.099	1.095	4.273	−3.435	0.433	0.716	H3	0.101	−0.276	−3.150	0.673	0.848	−0.850
C4	0.190	0.483	−2.430	−4.827	0.785	0.577	H4	0.061	0.045	0.760	1.568	0.559	0.211
O4	−0.650	−0.544	2.397	2.253	−0.706	−0.764	H5	0.400	0.598	−0.340	−0.583	0.582	0.674
C5	0.190	0.646	−6.136	−0.720	−0.112	0.850	H6	0.061	−0.133	2.029	−0.155	0.248	0.031
O5	−0.650	−0.769	−0.741	0.748	−0.796	−0.483	H7	0.400	0.663	1.503	1.273	0.830	0.366
C6	−0.099	1.099	3.019	2.946	−0.498	−0.388	H8	0.101	−0.270	2.005	−2.289	0.850	−0.850
N7	−0.250	−1.197	1.870	1.275	−0.850	−0.396	H9	0.051	0.000	−0.592	0.107	0.418	−0.850
C7	0.091	0.394	0.746	−1.794	0.236	0.492	H10	0.051	−0.160	0.604	0.595	0.850	−0.850
C20	0.001	−0.541	−0.191	3.592	−0.850	0.844	H11	0.131	−0.016	−1.944	1.854	0.196	−0.850
C21	−0.129	0.242	5.369	−6.820	−0.848	0.823	H12	0.131	0.045	−1.224	−2.102	0.399	−0.179
C22	−0.129	−0.137	0.967	9.451	−0.662	−0.026	H13	0.131	0.203	−0.984	−0.023	0.012	−0.083
C23	0.000	0.272	−2.577	−6.835	0.847	0.654	H14	0.131	0.266	1.707	−1.018	0.075	0.174
C24	−0.129	−0.295	5.402	0.350	−0.137	−0.530	H15	0.131	−0.293	−3.158	1.617	0.256	−0.556
C25	−0.129	−0.675	−6.798	4.066	−0.029	−0.474	H16	0.131	−0.066	0.064	−1.183	−0.157	−0.094
C26	−0.129	0.717	6.173	−6.145	−0.483	0.791	H17	0.131	0.289	0.940	−0.671	0.818	0.747
C27	−0.129	−0.200	0.612	6.736	0.257	−0.705	H18	0.051	0.112	0.165	−0.157	−0.101	0.850
C28	0.000	−0.191	−4.932	−0.662	−0.312	0.081	H19	0.051	0.282	0.539	1.128	0.850	−0.318
C29	−0.129	0.281	2.136	−1.460	−0.288	0.085	H20	0.131	−0.021	2.987	0.254	0.056	0.785
C31	−0.099	−0.622	−3.569	−0.673	−0.744	0.009	H21	0.131	−0.016	0.082	0.683	0.381	−0.253
C32	0.001	−0.077	−1.849	2.175	−0.773	−0.116	H22	0.131	−0.374	−0.647	−0.261	0.363	0.002
C33	−0.129	0.123	−0.799	−0.956	−0.410	−0.258	H23	0.131	−0.025	0.585	−0.896	0.111	0.177
C34	−0.129	−0.173	−1.981	−1.800	−0.407	0.023	H24	0.131	0.183	−0.592	0.657	0.512	−0.107
C35	−0.129	0.445	2.466	1.128	−0.048	−0.587	H25	0.051	0.119	−0.213	−0.839	0.714	0.850
C36	−0.129	−0.484	−4.401	1.563	−0.355	−0.849	H26	0.051	0.256	2.569	1.315	0.727	−0.445
C37	−0.129	0.218	5.025	−1.393	−0.324	0.452	H27	0.131	0.364	−0.857	−0.555	0.648	0.156
C61	−0.099	−0.709	−5.480	−1.806	−0.849	0.474	H28	0.131	−0.207	0.217	−1.438	0.099	−0.013
C62	0.001	0.203	1.399	−0.584	−0.845	0.097	H29	0.131	−0.472	4.420	−0.550	0.381	−0.011
C63	−0.129	−0.304	2.560	1.308	−0.034	0.306	H30	0.131	0.014	−2.298	0.239	0.515	−0.072
C64	−0.129	−0.344	−1.554	2.325	−0.413	−0.845	H31	0.131	−0.048	3.324	−0.220	−0.849	0.107
C65	−0.129	0.643	−6.252	−0.464	−0.580	−0.503	H32	0.131	0.332	1.252	−1.748	0.695	0.499
C66	−0.129	−0.232	6.064	−0.743	0.222	−0.096	H33	0.131	−0.128	−0.384	−1.502	0.241	−0.380
C67	−0.129	0.107	−5.391	1.244	−0.331	0.072	H34	0.131	−0.368	−0.579	−3.557	0.748	0.251
C70	0.001	−0.534	−3.400	0.430	−0.407	0.850	H35	0.131	0.530	−0.427	2.138	0.328	0.456
C71	−0.129	0.315	−1.799	2.497	−0.850	0.727	H36	0.131	0.131	0.190	0.787	−0.005	−0.051
C72	0.000	−0.220	−0.896	−3.430	−0.439	0.490	H37	0.131	0.051	−2.334	0.262	0.074	−0.126
C73	−0.129	−0.186	2.585	3.400	0.168	−0.039	H38	0.131	−0.019	−0.577	−0.455	0.557	−0.650
C74	−0.129	0.999	−2.084	4.949	−0.692	−0.271	ΔG^{rigid}	−15.1	−113.	826.	337.	126.	111.
C75	−0.129	−1.184	2.762	−5.547	−0.377	−0.665	ΔG^{Flex1}	−17.9	−113.	−87.1	176.	−325.	−343.
C76	−0.129	−0.008	−2.024	−0.171	−0.027	−0.062	ΔG^{Flex2}	−14.6	−114.	337.	−87.5	−101	−225.
C77	0.000	0.190	−1.792	1.870	0.338	−0.420							

[a] CHARMM: CHARMM charges of the ligand XK263. Optimized (rigid): charges optimized by solving eq 15 when the ligand is assumed to have the same conformation in free and bound states. Stationary point (Flex 1): charges obtained by solving eq 15 when the ligand is assumed to adopt conformation Flex 1 in the free state. Stationary point (Flex 2): same for the second alternative conformation of the free ligand, Flex 2. Minimized 1 (Flex 1): charges obtained by PRAXIS minimization of $\Delta G_{\text{ll,lp}}$, when the free ligand is in conformation Flex 1; added energy terms restrain net charge of the ligand to 0 and the charge of each atom to $|q_i| \leq 0.85$. Minimized 2 (Flex 1): a second set of charges obtained by the same method but starting from a different initial guess at the charges. $\Delta G_{\text{ll,lp}}^{\text{rigid}}$: energy obtained by substituting each set of charges into eq 17 when the **A** matrix is based upon assumption that the ligand is rigid, so the free conformation is the same as the bound conformation. $\Delta G_{\text{ll,lp}}^{\text{Flex1}}$: same, when the **A** matrix is computed assuming the free ligand is in conformation Flex 1. $\Delta G_{\text{ll,lp}}^{\text{Flex2}}$: same, when the **A** matrix is computed assuming the free ligand is in conformation Flex 2.

this concept by showing how the electrostatic energy of binding $\Delta G_{\text{ll,lp}}$ varies when all atoms of XK263 are changed gradually from their CHARMM charges to their optimal values (blue) and also when the charge of each individual atom is changed to its value in the optimal set of charges (black and red lines), while the other charges are held fixed at their CHARMM values. The parabolic shapes of these graphs are as expected from the theory (see section 2.2). As

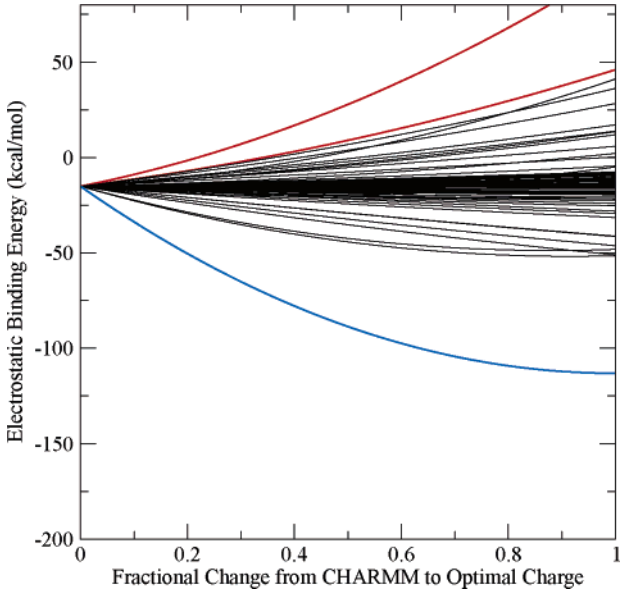


Figure 4. Electrostatic part of binding energy for rigid ligand as a function of the fractional shift from the initial CHARMM charges of XK263 (columns 2 and 9 of Table 2) toward charges optimized for the rigid ligand (columns 3 and 10 of Table 2). **Red**: energy change when each nitrogen atom's charge is varied, with all other atomic charge held fixed. **Black**: same as black graph, for the other 82 atoms of XK263. **Blue**: consequences of changing all ligand charges simultaneously to their optimal values.

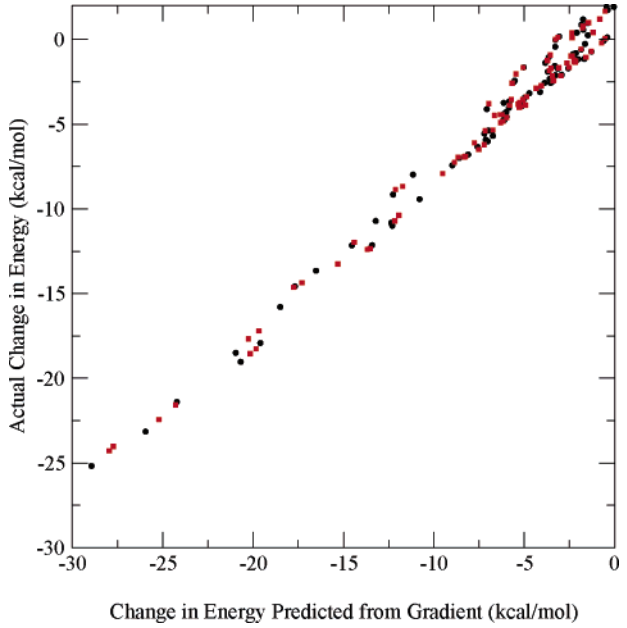


Figure 5. Accuracy of predictions from sensitivity analysis when the ligand is assumed flexible, shown as scatter plots of the change in electrostatic energy computed with the full parabolic energy surface versus the change predicted by sensitivity analysis, for charge changes of 0.2 au. Results are shown in black and red for two alternative free conformations, Flex 1 and Flex 2.

noted in the previous paragraph, changing all the charges to their optimal values produces a very large improvement in the electrostatic energy change upon binding. However, adjusting the charges of an individual atom toward its optimal value does not always improve the electrostatic binding energy. In fact, some changes markedly increase the energy and thus oppose binding.

For example, the two red lines correspond to the nitrogen atoms of XK263, which are situated roughly 5.5 Å from the aspartate groups at the bottom of the active site and roughly 4.2 Å from the amide hydrogens of the flaps at the top of the active site. The CHARMM force field assigns both nitrogens charges of 0.25 au, but the optimal charges are −1.7 and −1.2 au. (See Table 2.) It was initially surprising that optimization directs both atoms to be considerably more negative than their CHARMM values, even though the nearest protein ions are the negative aspartates, and both nitrogens as a consequence have strongly negative energy derivatives $\partial\Delta G/\partial q_i$. (See Table 1.) The explanation appears to be that charge optimization causes atoms closer to the aspartates, and with even more strongly negative derivatives than the nitrogens, to gain substantial positive charge, and these new positive charges effectively shield the nitrogens from the aspartates. In particular, atoms C3 and C6 (Figure 1) change from 0.099 au to 1.1 au, while atoms C4 and C5 also become significantly more positive. (See Table 2.) The resulting large increase in the positive charge situated between the aspartate groups and the nitrogens tends to drive the charges of the nitrogens in the positive direction in the final optimized charge set. This example explains why shifting a subset of the ligand's charges toward their optimal values may not improve the calculated binding affinity, in accord with previous observations.[9]

4.2. Flexible Ligand. Two alternative free conformations of the free ligand XK263 were generated by a simple molecular dynamics approach, as described in Methods, and were used separately as a basis for examining the consequences of ligand flexibility for sensitivity analysis and charge optimization.

4.2.1. Free Energy Derivatives. Using different conformations for the free ligand produces little change in the nature of the free energy derivative results. In particular, the changes in energy due to 0.2 au changes in the charges of single atoms are still well predicted by the derivatives, as shown in Figure 5. Moreover, the energy derivatives themselves are quite similar to those obtained under the assumption of a rigid ligand, as shown in Figure 6.

4.2.2. Charge Optimization. The consequences of ligand flexibility for charge optimization are more complex. First, as shown in Figure 3, the top left $N \times N$ submatrix of **A** matrix now possesses roughly 25 negative eigenvalues when either of the new free conformations is considered (red, green graphs). This implies that the stationary point of the energy surface in the absence of any constraints is a multidimensional saddle point. Thus, there is no optimal set of charges, at least in the case where the atomic charges are completely unconstrained. Instead the electrostatic binding energy can decrease without bound as a function of the charges assigned to the ligand.

Nonetheless, the method of Lagrangian multipliers can still be used to obtain a set of charges which is a stationary point of the energy on the multidimensional hyperplane defined by the constraint $\sum_i^N q_i = Q$. Columns 4, 5, 11, and 12 of

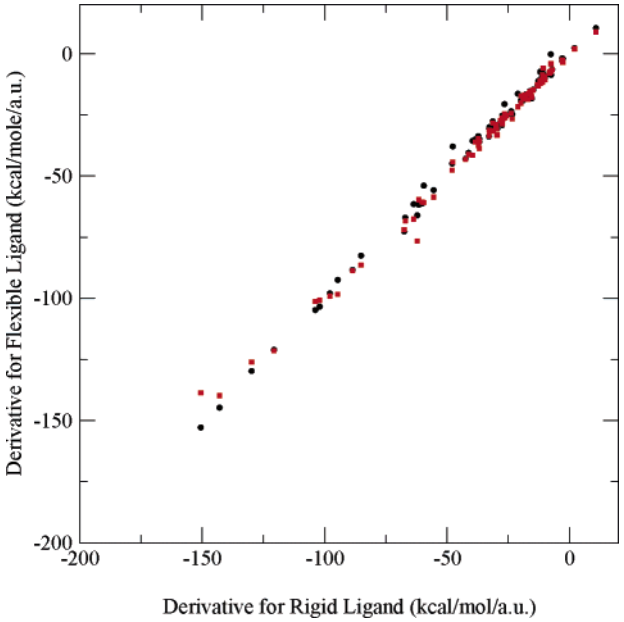


Figure 6. Comparison of free energy derivatives from the sensitivity analysis for two alternative conformations of free ligand, Flex 1 (black) and Flex 2 (red), with derivatives when the ligand is assumed rigid.

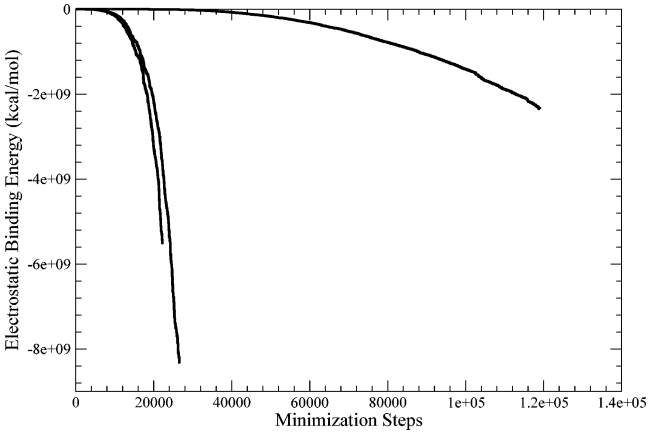


Figure 7. Electrostatic energy as a function of the number of PRAXIS minimization steps, when the free ligand conformation is taken to be different from the bound conformation (Flex 1) and the total ligand charge is restrained to zero. Results are shown for 3 different initial guesses of the atomic charges. The rightmost curve was obtained with a stronger restraining potential on the total ligand charge than the other two curves.

Table 2 present these stationary charges for the two free ligand conformations; they may be compared with those obtained with the assumption of a rigid ligand (columns 3 and 10). The mean absolute values of the stationary charges obtained with the two alternative free conformations are substantially larger, about 2 au, than the optimal charges based upon the assumption of a rigid ligand, about 0.4 au. There is no discernible correlation among the various sets of charges. These results show that changing the free conformation assumed for the ligand can lead to markedly different charge sets when the method of Lagrangian multipliers is used.

The stationary charges for each free ligand conformation can still lead to markedly improved binding energies, as shown at the foot of Table 2. Thus, when the stationary charges for the first flexible conformation are used to compute the binding energy based upon this free conformation, the result is −87.1 kcal/mol; the corresponding energy for the second conformation is similar, at −87.5 kcal/mol. Oddly, however, when these charge sets are used to compute energies under the assumption of alternative free conformations, very poor energies are obtained. For example, when the stationary charges obtained with the first flexible conformation (Table 2, columns 4 and 11) are used with the assumption of a rigid ligand, the energy becomes 826. kcal/mol; and when the same charges are used to compute the binding energy with the free conformation set to the second flexible conformation, the energy becomes 337. kcal/mol.

In contrast, when charges are optimized under the assumption of a rigid ligand (columns 3 and 10), the resulting energies are remarkably stable and favorable at about −113. kcal/mol across all choices of the free conformation. Interestingly, the CHARMM charges also give rather stable energies across all three free conformations, with values of −15 to −18 kcal/mol. In summary, the stationary charges obtained with the two alternate free conformations yield energies that depend strongly upon conformation. In contrast, optimal charges based upon the assumption of a rigid ligand provide a larger improvement in the binding energy, and this improvement is far less sensitive to the choice of free conformation.

This analysis has relied so far on Lagrangian multipliers to find ligand charges that optimize binding energy subject to the constraint on total charge. However, as noted above, it is possible that the binding energy has no lower bound, even when the total charge is constrained, because the top left $N \times N$ submatrix of **A** matrix possesses negative eigenvalues. This possibility is tested here by using a numerical algorithm (PRAXIS;[33] see Methods) to seek charge sets that sum to zero and yield highly favorable binding free energies. Figure 7 shows the results of three such minimizations started with different initial guesses for the atomic charges; one applies a much stronger restraining potential to the net charge of the ligand. The electrostatic binding energies are found to decline precipitously, reaching values as low as -8×10^{-9} kcal/mol with no sign of reaching an asymptote. It is important to note that the total charge was successfully locked near zero: the net charge of the ligand at the end of the three minimizations graphed in Figure 7 are 0.02, 0.0009, and 0.00007 au. However, the atomic charges obtained from these minimizations are completely unrealistic, with values in the hundreds and thousands of atomic units (data not shown). For comparison, Figure 8 shows corresponding convergence plots when the ligand is treated as rigid. All three plots converge toward −113 kcal/mol, the value obtained by the method of Lagrangian multipliers. (See ΔG^{rigid}, columns 3 and 10, in Table 2.) This expected result supports the validity of the numerical methods used in the present study.

The present analysis proves that the stationary point provided by the Lagrangian method is not a true minimum and is instead a saddle point. More importantly, they show

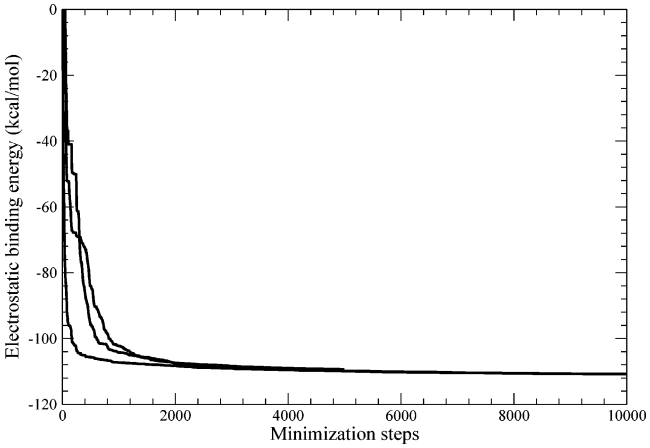


Figure 8. Electrostatic energy as a function of the number of PRAXIS minimization steps when the free ligand conformation is taken to be the same as the bound conformation and the total ligand charge is restrained to zero. Results for 3 different initial guesses of the atomic charges are shown.

that the charge optimization problem need not possess a well-defined solution when the conformation of the ligand is considered to change upon binding.

It is of interest to repeat the minimizations, now applying additional energy restraints that limit the absolute value of the charge of each atom to < 0.85 au, much as done in prior applications of charge optimization that treat the ligand as rigid.[6] These calculations yield convergent energies (data not shown), but the results vary from one minimization to another, depending upon the initial guess for the atomic charges. Two of the resulting charge sets, derived for the same conformation of the free ligand, are shown in columns 6, 7, 13, and 14 of Table 2. Both charge sets yield extremely favorable binding energies, −325 and −343 kcal/mol, when the free ligand is assumed to adopt the conformation for which the charges are optimized. These two charge sets also yield very low energies when the other alternative free conformation is assumed (−101 kcal/mol, −225 kcal/mol), but both are highly unfavorable when the free ligand is assumed to remain in the bound conformation (126 kcal/mol, 111 kcal/mol).

Remarkably, these charge sets can yield such large and negative values of $\Delta G_{ll,lp}$ that the total electrostatic energy of binding is found to be favorable even when the 98.4 kcal/mol penalty for desolvation of the protein is accounted for and when moreover all the electrostatic interactions between the ligand and the protein are artificially neglected. For example, for the second charge set (columns 7 and 14 in the table), the total electrostatic energy of binding is computed to be −31.6 kcal/mol under these assumptions.

This result at first appears necessarily incorrect on physical grounds, since it says that a favorable binding energy is obtained in the absence of any attractive forces between the ligand and the protein. However, the result is correct and is explained by the fact that the charge set in question produces intramolecular Coulombic interactions that powerfully stabilize the bound conformation of the ligand relative to the alternative free conformation, by about −600 kcal/mol, and this stabilization is only partly compensated by a desolvation energy difference of about +150 kcal/mol. As shown in Table 3, similar results apply to the other charge set generated by energy minimization but not to CHARMM charges or the charges generated by Lagrangian optimization when the ligand is assumed to be rigid. Thus, the minimization algorithm finds charges that massively stabilize the bound conformation relative to the alternative free conformation and thus appear to drive binding. It is important to emphasize that a molecule with such charges presumably would not in reality adopt the assumed free conformation but would instead be strongly preorganized into the bound conformation. As a consequence, most or all of the predicted electrostatic contribution to binding would be removed. This type of result is not obtained when when charges are optimized under the usual assumption that the free ligand remains in the bound conformation.

Table 3. Change in Electrostatic Energies (kcal/mol) When the Free Ligand Is Changed from the Flex 1 Conformation to the Bound Conformation (i.e., LigandConf^0 → LigandConf^1), Computed with Various Charge Sets[a]

	charge set			
	CHARMM	Rigid	Flex 1 Min 1	Flex 1 Min 2
Coulombic	−1.7	−12	−603	−607
Solvation (LPB)	−1	12	154	154
Total	−2.7	0	−449	−453

[a] Coulombic energies omit interactions between atoms in 1−2 and 1−3 bonded relationships. Solvation energies are computed with the finite difference, linearized Poisson−Boltzmann method, as described in the text. The total energy is the sum of the Coulombic and solvation terms. CHARMM: unoptimized CHARMM charges (columns 2 and 9 of Table 2). Rigid: charges optimized with the Lagrangian method under the assumption of a rigid ligand (columns 3 and 10 of Table 2). Flex 1 Min 1: the first set of charges adjusted with the PRAXIS algorithm to minimize binding energy when the ligand is assumed to adopt the Flex 1 conformation in the free state, with the total ligand charge restrained to 0 and the absolute values of individual charges restrained ≤ 0.85 au (columns 6 and 13, Table 2). Flex 1 Min 2: the second set of charges adjusted with the PRAXIS algorithm to minimize binding energy when the ligand is assumed to adopt the Flex 1 conformation in the free state, with the total ligand charge restrained to 0 and the absolute values of individual charges restrained ≤ 0.85 au (columns 7 and 14, Table 2).

5. Discussion

Sensitivity analysis is quite accurate for many charge changes large enough to be of interest in lead optimization (±0.5 au), even though it is a linear approximation to an energy function with parabolic curvature. However, its accuracy diminishes significantly for charge changes of ±1 au. Interestingly, accuracy is not degraded when the ligand's conformation is considered to change on binding. Indeed, the energy derivatives are surprisingly insensitive to the assumed conformation of the free ligand. This may result from the fact that the electrostatic potentials at the atoms of the free ligand are dominated by solvation terms which do not depend strongly upon conformation, rather than by conformation-dependent interatomic interactions. Sensitivity analysis is computationally efficient because it requires only one FDPB calculation for each conformation of the ligand. Even greater speed could be be achieved by use of general-

ized Born type models (see, e.g., refs 35−40). Given its efficiency, and the limitations of the linear approximation which forms its basis, sensitivity analysis might be best deployed iteratively. That is, derivatives can be computed and used to guide a first chemical change. Derivatives could then be recomputed for the revised compound to guide a second change and so on. It would furthermore be possible to recompute the conformational preferences of the free ligand after each change and thereby incorporate full-fledged Boltzmann averages of the potentials as derivatives, via eq 7. Finally, it is worth noting that the present generalization of sensitivity analysis is not limited to analyzing the sensitivity of binding free energy to ligand charges but has a much wider range of potential applications. For example, it could be used to examine the sensitivity of protein folding to the strength of an energy term controlling dihedral rotation.

The properties of charge-optimization are more complex, even when the ligand is assumed to be rigid. One important observation is that optimization of *all* ligand charges simultaneously does not appear to be a reliable means of identifying changes in *part* of the ligand that will increase affinity, as previously noted.[9] This issue, illustrated by the case of the urea nitrogens of XK263 in section 4.1.2, could be addressed by applying the optimization formalism only to the part of the ligand that is a candidate for chemical modification, while holding the rest of the charges constant.[9] Mathematically, this would involve merely deleting the rows and columns of the *A* matrix corresponding to the charges which are to be held constant and then applying the method of Lagrangian multipliers as usual. Charge-optimization also has traditionally been rather time-consuming, because it has required at least one FDPB calculation for each atom whose charge is to optimized. Recent methodological advances address this problem, however.[10,11] Charge-optimization, like sensitivity analysis, could be markedly accelerated by replacing FDPB calculations with faster generalized Born calculations.

When the ligand changes conformation upon binding, the Lagrangian charge-optimization formalism may not yield charges that are actually optimal. Instead the stationary point it provides can be a saddle, as suggested by the existence of negative eigenvalues for the upper left $N \times N$ submatrix of the **A** matrix and supported by minimization convergence plots with profoundly negative energies and no evidence of approaching an asymptote. In such cases, charges can be found that lower the nominal binding energy without limit, unless further restraints are applied. Interestingly, charges adjusted to yield a strong nominal binding energy in these circumstances do not driving binding as such but rather a conformational change of the ligand from the free to the bound conformation. There is no guarantee that these charges will actually drive binding. Instead, by strongly stabilizing the bound conformation of the ligand in the free state, these charges violate the assumption of a different free conformation used in generating the charges. More generally, the present analysis highlights the fact that a purely electrostatic analysis cannot yield truly optimal charges, because the best charges depend on the conformational preferences of the ligand which, in turn, are influenced by nonelectrostatic contributions to the energy.

From a practical standpoint, however, if one artificially assumes that the free conformation of the ligand is the same as the bound conformation, as normally done when charge-optimization is employed, the formalism yields charges that robustly improve binding across the different free conformations considered here, as shown at the foot of Table 2. Thus, the current practice of assuming a rigid ligand should often be effective.

In summary, the present results indicate that both sensitivity analysis and charge-optimization can help guide the conversion of a lead compound into a high affinity drug candidate. However, it is important to keep in mind that a predicted improvement in the electrostatic part of the binding energy obtained by applying these methods may not be expressed exactly in the standard free energy of binding because of nonelectrostatic factors. For one thing, it is never possible in reality to change atomic charges without changing other atomic properties, such as atomic radii. In addition, charge changes may produce unforeseen changes in the degree of preorganization of the ligand.

Acknowledgment. This publication was made possible by Grant Number GM61300 from the National Institute of General Medical Sciences (NIGMS) of the NIH. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIGMS. The author thanks the Netlib and LAPACK groups for making the PRAXIS, dsyev, and dgesv, routines available; Mr. Himan Mookherjee for discussions; and Dr. Michael J. Potter and the anonymous referees for their valuable comments.

Supporting Information Available: Three-dimensional coordinates, in PDB format, of the three conformations of XK263 used in the present study: LigandConformation[0] (XK263_docked.pdb), LigandConformation[1] (XK263_md-flex1.pdb), and LigandConformation[2] (XK263_md-flex2.pdb). This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Gilson, M. K.; Honig, B. *Proc. Natl. Acad. Sci. U.S.A.* **1989,** *86*, 1524−1528.

(2) Hendsch, Z. S.; Tidor, B. *Protein Sci.* **1994**, *3*, 211−226.

(3) Lee, L. P.; Tidor, B. *J. Chem. Phys.* **1997**, *106*, 8681−8690.

(4) Kangas, E.; Tidor, B. *J. Chem. Phy.* **1998**, *109*, 7522−7545.

(5) Sulea, T.; Purisima, E. O. *J. Phys. Chem. B* **2001**, *105*, 889−899.

(6) Kangas, E.; Tidor, B. *J. Phys. Chem. B* **2001**, *105*, 880−888.

(7) Sulea, T.; Purisima, E. O. *Biophys. J.* **2003**, *84*, 2883−2896.

(8) Mandal, A.; Hilvert, D. *J. Am. Chem. Soc.* **2003**, *125*, 5598−5599.

(9) Sims, P. A.; Wong, C. F.; McCammon, J. A. *J. Comput. Chem.* **2004**, *25*, 1416−1429.

(10) Bardhan, J. P.; Lee, J. H.; Kuo, S. S.; Tidor, M. D. A. B.; White, J. K. Fast methods for biomolecule charge optimization. In *Technical proceedings of the 2003 Nanotechnology Conference and Trade Show*; Nanoscience and Technology Institute: Cambridge, MA, 2003.

(11) Bardhan, J. P.; Lee, J. H.; Altman, M. D.; Leyffer, S.; Benson, S.; Tidor, B.; White, J. K. Biomolecule electrostatic optimization with an implicit Hessian. In *Technical proceedings of the 2004 Nanotechnology Conference and Trade Show*; Nanoscience and Technology Institute: Cambridge, MA, 2004.

(12) Zhang, H.; Wong, C. F.; Thacher, T.; Rabitz, H. *Proteins: Struct., Funct., Genet.* **1995**, *23*, 218−232.

(13) Wong, C. F.; Thacher, T.; Rabitz, H. Sensitivity analysis in biomolecular simulation. In *Rev. Comput. Chem*.; Wiley-VCH: New York, 1998.

(14) Sims, P. A.; Wong, C. F.; McCammon, J. A. *J. Med. Chem.* **2003**, *46*, 3314−3325.

(15) Gilson, M. K.; Given, J. A.; Bush, B. L.; McCammon, J. A. *Biophys. J.* **1997**, *72*, 1047−1069.

(16) Hill, T. L. *Statistical Mechanics. Principles and selected applications;* McGraw-Hill: New York, 1956.

(17) Cieplak, P.; Pearlman, D. A.; Kollman, P. A. *J. Chem. Phys.* **1994**, *101*, 627−633.

(18) Gilson, M. K.; Rashin, A. A.; Fine, R.; Honig, B. *J. Mol. Biol.* **1985**, *183*, 503−516.

(19) Arfken, G. *Mathematical Methods for Physicists*, 3rd ed.; Academic Press: Orlando, FL, 1985.

(20) Lam, P. Y. S.; Jadhav, P. K.; Eyermann, C. J.; Hodge, C. N.; Ru, Y.; Bacheler, L. T.; Meek, J. L.; Otto, M. J.; Rayner, M. M.; Wong, Y. N.; Chang, C.; Weber, P. C.; Jackson, D. A.; Sharpe, T. R.; Erickson-Viitanen, S. *Science* **1994**, *263*, 380−384.

(21) Lam, P. Y.; Jadhav, P. K.; Eyermann, C. J.; Hodge, C. N.; Ru, Y.; Bacheler, L. T.; Meek, J. L.; Otto, M. J.; Rayner, M. M.; et al., Y. N. W. *Science* **1994**, *263*, 380−384.

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

(23) Accelrys, Inc. San Diego, CA,.

(24) David, L.; Luo, R.; Gilson, M. K. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 157−171.

(25) Kairys, V.; Gilson, M. K. *J. Comput. Chem.* **2002**, *23*, 1656−1670.

(26) van Gunsteren, W. F.; Berendsen, H. J. C. *Mol. Simul.* **1988**, *1*, 173−185.

(27) Davis, M. E.; Madura, J. D.; Luty, B. A.; McCammon, J. A. *Comput. Phys. Commun.* **1991**, *62*, 187−197.

(28) Richards, F. M. *Annu. Rev. Biophys. Bioeng.* **1977**, *6*, 151−176.

(29) Warwicker, J.; Watson, H. C. *J. Mol. Biol.* **1982**, *157*, 671−679.

(30) Klapper, I.; Hagstrom, R.; Fine, R.; Sharp, K.; Honig, B. *Proteins: Struct., Funct., Genet.* **1986**, *1*, 47−79.

(31) Gilson, M. K.; Sharp, K. A.; Honig, B. H. *J. Comput. Chem.* **1988**, *9*, 327−335.

(32) Anderson, E.; Bai, Z.; Bischof, C.; Blackford, S.; Demmel, J.; Dongarra, J.; Du Croz, J.; Greenbaum, A.; Hammarling, S.; McKenney, A.; Sorensen, D. *LAPACK Users' Guide;* Society for Industrial and Applied Mathematics: Philadelphia, PA, 3rd ed.; 1999.

(33) http://www.netlib.org/opt/praxis.

(34) http://www.netlib.org.

(35) Qiu, D.; Shenkin, P. S.; Hollinger, F. P.; Still, W. C. *J. Phys. Chem.* **1997**, *101*, 3005−3014.

(36) Ghosh, A.; Rapp, C. S.; Friesner, R. A. *J. Phys. Chem. B* **1998**, *102*, 10983−10990.

(37) Jayaram, B.; Liu, Y.; Beveridge, D. L. *J. Chem. Phys.* **1998**, *109*, 1465−1471.

(38) Dominy, B. N.; Brooks, C. L., III. *J. Phys. Chem. B* **1999**, *103*, 3765−3773.

(39) Lee, M. S.; Salsbury, F. R., Jr.; Charles L.; Brooks, I.

(40) Im, W.; Lee, M. S.; Brooks, C. L., III. *J. Comput. Chem.* **2003**, *24*, 1691−1702.

CT050226Y

JCTC Journal of Chemical Theory and Computation

An Atoms in Molecules Study of the Halogen Resonance Effect

Norberto Castillo and Russell J. Boyd*

Department of Chemistry, Dalhousie University, Halifax, N.S., Canada B3H 4J3

Received September 22, 2005

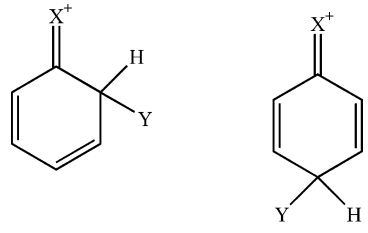
Abstract: We report a detailed study by means of the theory of atoms in molecules (AIM) of the resonance effect exhibited in systems where a halogen is adjacent to a carbon−carbon double bond. Moreover, we have carried out a comparable study of the respective saturated halohydrocarbons and hydrocarbons, as well as the related unsaturated hydrocarbons. The valence shell charge concentration (VSCC) of the atoms in systems that exhibit the halogen resonance effect is considerably different from that of the systems where only the electron withdrawing inductive effect is present. Our analysis of the bonded maximum charge concentration and the electronic properties at the bond critical points clearly indicate that the carbon−carbon double bond is strongly distorted as a result of the halogen resonance effect. Population analyses show that the halogen resonance effect is a donor effect, but the opposing electron-withdrawing inductive effect is stronger. Moreover, the analysis in terms of link points of the VSCCs of the carbons accounts for the observed position-dependence of electrophilic aromatic substitution in α- and β-halonaphthalenes.

1. Introduction

The electronic structure of certain classes of molecular species cannot be adequately described by a single Lewis structure. In some cases, the actual electronic structure is a weighted average of two or more Lewis structures, called resonance structures, and the molecule is known as a resonance hybrid. The concept of resonance is especially useful for systems containing delocalized electrons and has been used to explain many phenomena in chemistry including several types of reactions and the stability and physical properties of compounds.

The experimental observation that electrophilic substitution at the ortho and para positions of halobenzenes (C_6H_5X) is more facile than at the meta position is readily rationalized by a halogen resonance effect. Thus, interaction of a halogen lone pair with the p atomic orbitals that form the delocalized system of π bonds leads to the halonium ion structures shown in Chart 1. It is impossible to draw an equivalent resonance structure for the intermediate formed by electrophilic substitution at the meta position of a halobenzene.

Chart 1. Resonance Structures for the Intermediates Formed by Electrophilic Aromatic Substitution of Y^+ at the Ortho (a) and Para (b) Positions of a Halobenzene



(a) (b)

Support for the standard interpretation of the experimental results is provided by molecular orbital (MO) calculations,[1] which clearly show a large contribution of the halogen to the π-bonding MOs. A similar effect is not observed in haloaliphatic compounds. In this paper, we report the first analysis of the halogen resonance effect by use of the theory of atoms in molecules (AIM).

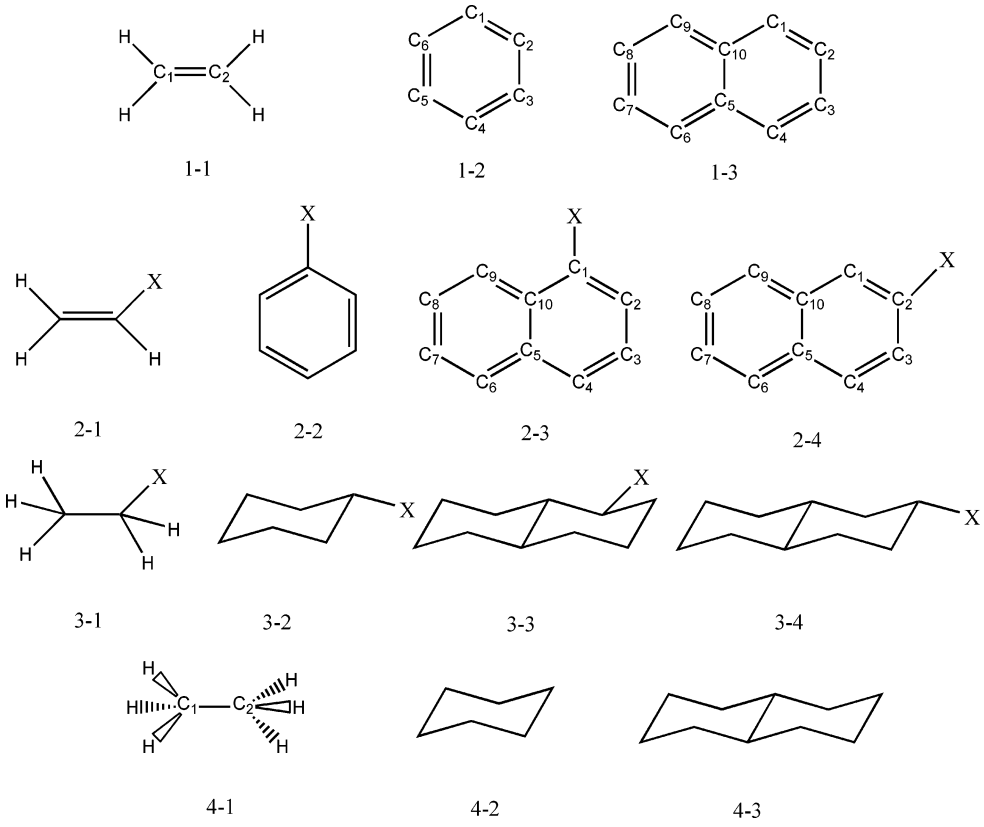
2. The Theory of Atoms in Molecules and Resonance

The AIM theory uses well-defined quantities derived from the electron density to provide valuable insight into the

* Corresponding author tel.: (902) 494-8883; fax: (902) 494-1310; e-mail: russell.boyd@dal.ca.

10.1021/ct050238j CCC: $33.50
Published on Web 12/23/2005

Chart 2. Chemical Structures of Compounds Included in This Study, Where X = F, Cl, and Br



electronic structures and properties of molecules.[2−7] Bader[8] has studied the resonance effect in terms of $F^{\alpha}(\Omega,\Omega')$ and $F^{\beta}(\Omega,\Omega')$, which are the delocalization functions of electrons with α and β spins, respectively, between the basins of two atoms, Ω and Ω′. The delocalization function for α spin for a Slater determinantal wave function is given by

$$F^{\alpha}(\Omega,\Omega') = -\sum_i \sum_j \int \mathrm{d}r_1 \int \mathrm{d}r_2 \{\phi_i^*(r_1)\phi_j(r_1)\phi_j^*(r_2)\phi_i(r_2)\} = -\sum_{ij} S_{ij}(\Omega)\, S_{ji}(\Omega') \quad (1)$$

where the ϕ's are the α spin−orbitals and $S_{ij}(\Omega)$ denotes the overlap of a pair of α spin−orbitals over Ω (basin). An equivalent expression holds for β spin. Essentially, these delocalization functions measure the sharing of electrons between two atoms. The relationship between the delocalization index and bond order in the characterization of a chemical bond has been discussed previously.[9−12]

Bader et al.[13] used $F^{\alpha}(\Omega,\Omega')$ to quantify the contribution of each resonance structure in acyclic and cyclic hydrocarbons as well as the effect of substituents on delocalization in aromatic systems. More recently, González and Mosquera[14] reported a similar study in pyrimidinic bases but in terms of the delocalization index [$\delta(\Omega,\Omega')$], which is a more general parameter containing both delocalization functions:

$$\delta(\Omega,\Omega') = 2|F^{\alpha}(\Omega,\Omega')| + 2|F^{\beta}(\Omega,\Omega')| \quad (2)$$

Several other authors have used AIM to study the resonance effect in many different types of systems. For example, Okulik et al. have used AIM parameters to study the "three-center two-electron bonds" exhibited by isobutonium[15] and *n*-butonium[16] cations. Grabowski[17,18] and Gilli et al.[19] have used AIM analysis in terms of the electron density and the Laplacian at the bond critical points to explore the resonance-assisted hydrogen bonds in malonaldehyde and ketohydrazone−azoenol systems, respectively. Also, Borbulevych et al.[20] used the AIM theory to analyze substituent effects in 4-nitroaniline derivatives.

In this paper, we present a detailed AIM study of systems where the halogen is adjacent to a carbon−carbon double bond. Moreover, we have carried out a comparable study of the respective saturated halohydrocarbons and hydrocarbons as well as the related unsaturated hydrocarbons. The molecules included in this study are shown in Chart 2, where X = F, Cl, and Br. The first series consists of ethene (1-1), benzene (1-2), and naphthalene (1-3). The second series consists of their four monohalo derivatives. The third series consists of the saturated analogues of series 2, while the fourth series consists of the unsubstituted parent compounds of the third series. The emphasis of our study is on bond critical points, ellipticities, Laplacian topology, delocalization indexes, and population analysis for the main atoms and bonds associated with the resonance effect.

A secondary purpose of this paper is to perform an AIM study of electrophilic aromatic substitution in the halonaphthalenes in order to complement Bader and Chang's earlier study of benzene.[21] We use an analysis of the link points of the carbon valence shell charge concentrations (VSCCs) of the ring to predict the directing and activating-deactivating effects of halogens in naphthalene.

3. Computational Details

All molecules were fully optimized at the B3LYP/6-311++G(d,p) level using the Gaussian 03 package.[22] The characterization of the bond and ring critical points as well as the maximum charge concentrations was carried out using the EXTREME program, while the atomic populations were performed by PROAIM. Both programs belong to the AIMPAC package.[23,24] The AIMDELOC program was used to obtain the delocalization indexes.[25]

4. Results and Discussion

4.1. Valence-Shell Characterization of Chlorine in Compound Series 2 And 3. The characterization of the valence shell was carried out in terms of the (3, −3) critical points of L ($L = -\nabla^2\rho$), which represent the bonded and nonbonded maximum charge concentrations in the VSCC of an atom in a molecule. The locations of the (3, −3) critical points of L provide theoretical support for the bonded and nonbonded electron pairs of the Lewis model.[26−28] Figure 1 illustrates the locations of the (3, −3) critical points in the valence shell of chlorine in compound series 2 and 3. The position of each (3, −3) critical point is indicated by a vector whose origin is at the chlorine nucleus. To illustrate the distortion of the VSCC in each case, Figure 2 shows the contour map of the Laplacian for the plane that contains the halogen and the two carbons of chloroethane and chloroethene.

Figure 1 clearly shows that the chlorine VSCC exhibits two nonbonded maxima in series 2 and three nonbonded maxima in series 3. Moreover, the missing nonbonded maxima in series two are positioned optimally to delocalize the π cloud of the carbon−carbon double bonds. Figure 1 also illustrates that the other two nonbonded maxima are in the σ plane of the carbon−carbon double bonds, which is a favorable location for the delocalization of the missing nonbonded maximum charge concentration into the π cloud. Furthermore, Figure 2 shows contour lines connecting the VSCC of chlorine with the VSCC of carbon in chloroethene, signifying a greater distortion of the VSCCs of chlorine and carbon in chloroethene than in chloroethane. This suggests a greater sharing of electrons in the chlorine−carbon bond of chloroethene than that in chloroethane and also suggests that the halogen resonance effect is a donor effect.

Table 1 describes the VSCC of chlorine in compound series 2 and series 3 in terms of several electronic properties. The radii, $-\nabla^2\rho$, and ρ at the bonded maximum charge concentration of the chlorine VSCC are larger in series 2 than in series 3. The same trend holds for the nonbonded maximum charge concentrations, with the exception of the radii, which are slightly larger in series 3 than in series 2. The radii of the nonbonded maximum charge concentrations are smaller in series 2 than series 3, whereas $-\nabla^2\rho$ and ρ continue being larger. The decrease of the angle from series 2 to series 3 indicates the change of the chlorine VSCC from trigonal planar to tetrahedral. All these results support the fact that one nonbonded maximum is delocalized into the π cloud of the double bonds in series 2.

4.2. Comparison of the Fluorine and Bromine Valence Shells with Chlorine in Compound Series 2 and Series 3. The bonded maximum charge concentrations of the VSCCs of fluorine and bromine in series 2 and series 3 were not found at the level of theory used in this study. The results are listed in Table 2. It is possible that the maxima would be found at higher levels of calculation since it is known that the use of a triplet-ζ basis set is the minimum requirement to obtain consistent and topologically stable graphs of the Laplacian.[29]

The nonbonded maxima charge concentrations for fluorine and bromine have similar characteristics to that of chlorine. For example, the VSCCs for the two atoms exhibit two

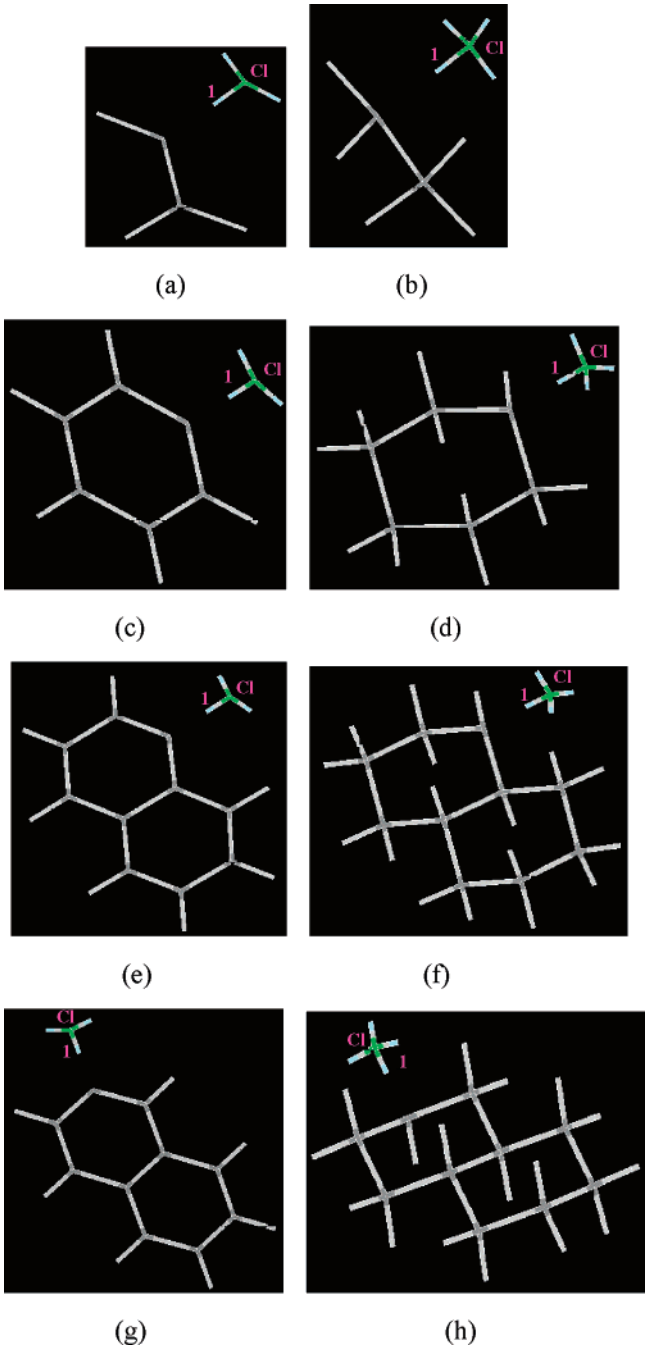


Figure 1. Location of the maximum charge concentrations of the VSCCs of chlorine in compound series 2 and series 3. In each molecule, the bonded maximum charge concentration is directed toward the carbon atom bonded to chlorine and is labeled by 1. Representations are as follows: (a) 2-1, (b) 3-1, (c) 2-2, (d) 3-2, (e) 2-3, (f) 3-3, (g) 2-4, and (h) 3-4.

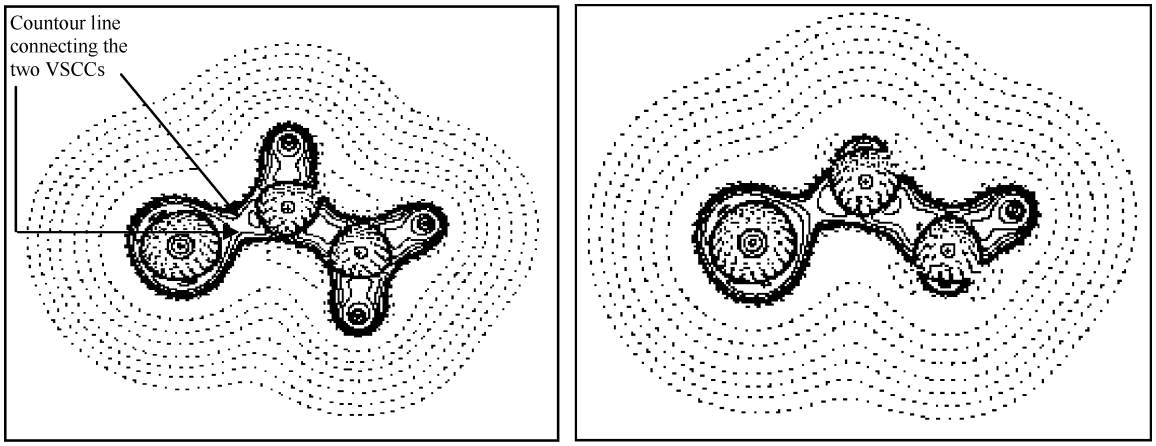


Figure 2. Contour map of the Laplacian of the electron density in the plane that contains the chlorine and the two carbons. (a) Chloroethene and (b) chloroethane. The chlorine nucleus is on the left side in both cases.

Table 1. Characterization of the Maximum Charge Concentrations in the VSCCs of Chlorine in Compound Series 2 and Series 3 in Terms of the Number of Bonded Maxima (# b), the Number of Nonbonded Maxima (# nb), Radius (r), $-\nabla^2\rho$, ρ, and Average Angles between Nonbonded Maxima[a]

	bonded maxima				nonbonded maxima				
molecule	# b	r	$-\nabla^2\rho \times 10^1$	$\rho \times 10^1$	# nb	r	$-\nabla^2\rho \times 10^1$	$\rho \times 10^1$	angle (nb−nb)
2-1	1	1.271	6.18	2.55	2	1.185	8.52	2.78	152.4
						1.185	8.50	2.78	
3-1	1	1.265	5.76	2.45	3	1.187	8.31	2.76	115.2
						1.187	8.31	2.76	
						1.187	8.30	2.76	
2-2	1	1.271	6.09	2.53	2	1.185	8.47	2.78	153.1
						1.185	8.47	2.78	
3-2	1	1.261	5.74	2.45	3	1.188	8.19	2.75	115.5
						1.188	8.21	2.75	
						1.188	8.21	2.75	
2-3	1	1.271	6.03	2.52	2	1.185	8.47	2.78	153.4
						1.185	8.49	2.78	
3-3	1	1.258	5.73	2.51	3	1.187	8.21	2.75	115.1
						1.188	8.19	2.75	
						1.187	8.21	2.75	
2-4	1	1.271	6.08	2.53	2	1.185	8.47	2.78	153.7
						1.185	8.47	2.78	
3-4	1	1.261	5.73	2.45	3	1.188	8.19	2.75	115.3
						1.188	8.21	2.75	
						1.187	8.21	2.75	

[a] Radius in angstroms, $-\nabla^2\rho$ and ρ in atomic units, and angles in degrees.

nonbonded maxima in series 2, whereas there are three nonbonded maxima in series 3. Also, the nonbonded maximum charge concentrations in series 2 are in a favorable location for the delocalization of the missing nonbonded maximum charge concentration into the π cloud. The radius of the nonbonded maximum charge concentration increases from series 2 to series 3, whereas $-\nabla^2\rho$, ρ, and the angles decrease. A decrease of almost 75% in the $-\nabla^2\rho$ value for the nonbonded maximum charge concentration in the bromine case is noteworthy. This result is expected because bromine is considerably larger than chlorine and fluorine, and any displacement of the same amount of electron charge involves larger volumes. Therefore, the concentration of ρ decreases more significantly than in the cases of chlorine and fluorine.

4.3. Characterization of the Bonded Maximum Charge Concentrations of the Carbons Connected to Halogens. As was shown above, the VSCCs of the halogens are significantly altered by the resonance effect. Hence, it can be expected that the VSCCs of the carbon atoms joined to the halogen will also exhibit detectable modifications. Therefore, we analyzed the VSCCs of these carbon atoms for series 2 and series 3 with the three different halogens. Table 3 provides the results for the bonded maximum charge concentration of the carbon bonded to the halogens. As can be seen in the table, the systems with the resonance effect

Table 2. Characterization of the Nonbonded Maximum Charge Concentrations in the VSCCs of Fluorine and Bromine in Series 2 and Series 3 in Terms of the Number of Nonbonded Maxima (# nb), Radius (r), $-\nabla^2\rho$, ρ, and Average Angles between Nonbonded Maxima[a]

Nonbonded Maxima											
	fluorine						bromine				
molecule	# nb	r	$-\nabla^2\rho\times10^2$	$\rho\times10^1$	angle (nb−nb)	molecule	# nb	r	$-\nabla^2\rho$	r	angle (nb−nb)
2-1	2	0.569	9.34	1.52	157.0	2-1	2	1.567	0.82	1.65	158.9
		0.569	9.31	1.52				1.568	0.82	1.65	
3-1	3	0.571	9.06	1.50	116.2	3-1	3	1.574	0.21	1.62	117.1
		0.571	9.06	1.50				1.574	0.19	1.62	
		0.571	9.07	1.50				1.574	0.19	1.62	
2-2	2	0.569	9.28	1.52	155.7	2-2	2	1.567	0.76	1.65	159.8
		0.569	9.28	1.51				1.567	0.76	1.65	
3-2	3	0.571	9.00	1.50	116.6	3-2	3	1.577	0.16	1.61	117.3
		0.571	8.98	1.50				1.576	0.03	1.61	
		0.571	9.00	1.50				1.576	0.03	1.61	
2-3	2	0.569	9.26	1.52	156.1	2-3	2	1.566	0.94	1.66	159.8
		0.569	9.26	1.52				1.566	0.77	1.65	
3-3	3	0.571	9.01	1.50	116.0	3-3	3	1.576	0.11	1.61	116.7
		0.571	8.96	1.50				1.578	0.29	1.60	
		0.571	8.98	1.50				1.575	0.09	1.62	
2-4	2	0.569	9.28	1.52	156.6	2-4	2	1.567	0.79	1.65	159.3
		0.569	9.28	1.52				1.567	0.77	1.65	
3-4	3	0.571	8.97	1.50	116.6	3-4	3	1.577	0.16	1.61	117.2
		0.571	9.00	1.50				1.576	0.03	1.61	
		0.571	9.00	1.50				1.576	0.03	1.61	

[a] Radius in angstroms, $-\nabla^2\rho$ and ρ in atomic units, and angles in degrees.

Table 3. Characterization of the Carbon−Halogen Bonded Maximum Charge Concentrations in the VSCC of the Carbon Connected to the Halogen (Fluorine, Chlorine, or Bromine) in Terms of Radius, $-\nabla^2\rho$, and ρ for Series 2 and Series 3[a]

fluorine				chlorine				bromine			
molecule	r	$-\nabla^2\rho\times10^1$	$\rho\times10^1$	molecule	r	$-\nabla^2\rho\times10^1$	$\rho\times10^1$	molecule	r	$-\nabla^2\rho\times10^1$	$\rho\times10^1$
2-1	1.043	3.47	2.63	2-1	1.035	4.99	2.13	2-1	1.027	4.81	2.02
3-1	1.053	2.64	2.29	3-1	1.054	3.84	1.89	3-1	1.049	3.53	1.79
2-2	1.044	3.51	2.62	2-2	1.034	4.95	2.11	2-2	1.026	4.80	2.01
3-2	1.052	2.45	2.23	3-2	1.054	3.61	1.85	3-2	1.049	3.33	1.75
2-3	1.045	3.43	2.61	2-3	1.035	4.84	2.10	2-3	1.027	4.68	1.99
3-3	1.054	2.37	2.21	3-3	1.055	3.53	1.83	3-3	1.050	3.20	1.73
2-4	1.044	3.51	2.62	2-4	1.034	4.94	2.11	2-4	1.026	4.80	2.01
3-4	1.053	2.44	2.23	3-4	1.055	3.59	1.84	3-4	1.050	3.31	1.75

[a] Radius in angstroms and $-\nabla^2\rho$ and ρ in atomic units.

(series 2) exhibit smaller radii and greater values of $-\nabla^2\rho$ and ρ at the maximum charge concentration of the carbon bonded to the halogen than those without the resonance effect (series 3). This fact can be explained by the delocalization of the nonbonded charge concentration in the carbon−halogen bond. The resonance effect plays a similar role in the unsaturated aliphatic halohydrocarbon to that of the aromatic halohydrocarbon. There are no appreciable differences with respect to the radius, $-\nabla^2\rho$, and ρ between the haloethene (2-1), the halobenzene (2-2), and the halonaphthalenes (2-3 and 2-4), respectively. Fluorine produces the smallest value of $-\nabla^2\rho$ and the largest value of ρ at the bonded maximum charge concentration of the carbon bonded to the halogen.

We analyzed the carbon connected to the halogen in series 2 and series 3 in terms of the bonded charge concentration contained in the carbon−carbon bond in order to investigate the halogen resonance effect on the carbon−carbon double bond in series 2. Also, we characterized the respective bonded charge concentrations of the carbons for the nonhalogen systems (series 1 and series 4) to make illustrative comparisons. Table 4 provides the results. The VSCC of the carbon connected to the halogen in series 2 and series 3 exhibits a bonded maximum charge concentration contained in the carbon−carbon bond with smaller radii than the nonhalogen systems (series 1 and series 4). The radii are even smaller in cases where the halogen resonance effect exists (series 2). Moreover, series 2 exhibits the greatest values of $-\nabla^2\rho$ and ρ at these bonded maximum charge concentrations. The values of $-\nabla^2\rho$ and ρ for nonhalogen unsaturated systems (series 1) are between the two types of halogen systems (series 2 and series 3). Fluorine produces smaller radii and larger values of $-\nabla^2\rho$ and ρ in its compounds than chlorine and bromine. Slight differences are

Table 4. Characterization of the Carbon−Carbon Bonded Maximum Charge Concentrations in the VSCCs of the Carbon Connected to the Halogen (Fluorine, Chlorine, or Bromine) in Terms of Radius, $-\nabla^2\rho$, and ρ for All Four Series[a]

fluorine				chlorine				bromine			
molecule	*r*	$-\nabla^2\rho$	$\rho \times 10^1$	molecule	*r*	$-\nabla^2\rho$	$\rho \times 10^1$	molecule	*r*	$-\nabla^2\rho$	$\rho \times 10^1$
1-1	0.987	1.14	3.53	1-1	0.987	1.14	3.53	1-1	0.987	1.14	3.53
2-1	0.958	1.28	3.71	2-1	0.968	1.21	3.63	2-1	0.970	1.19	3.61
3-1	0.966	1.03	2.94	3-1	0.970	0.98	2.88	3-1	0.971	0.97	2.88
4-1	0.996	0.82	2.68	4-1	0.996	0.82	2.68	4-1	0.996	0.82	2.68
1-2	0.983	1.06	3.25	1-2	0.983	1.06	3.25	1-2	0.983	1.06	3.25
2-2	0.958	1.20	3.43	2-2	0.965	1.13	3.34	2-2	0.967	1.11	3.32
3-2	0.962	1.05	2.96	3-2	0.966	1.00	2.91	3-2	0.967	0.99	2.90
4-2	0.991	0.85	2.72	4-2	0.991	0.85	2.72	4-2	0.991	0.85	2.72
1-3 (C_1)	0.982	1.09	3.34	1-3 (C_1)	0.982	1.09	3.34	1-3 (C_1)	0.982	1.09	3.34
2-3	0.957	1.23	3.52	2-3	0.965	1.15	3.43	2-3	0.967	1.13	3.41
3-3	0.962	1.05	2.96	3-3	0.965	1.00	2.92	3-3	0.966	0.99	2.90
4-3 (C_1)	0.991	0.85	2.72	4-3 (C_1)	0.991	0.85	2.72	4-3 (C_1)	0.991	0.85	2.72
1-3 (C_2)	0.983	1.08	3.34	1-3 (C_2)	0.983	1.08	3.34	1-3 (C_2)	0.983	1.08	3.34
2-4	0.957	1.23	3.51	2-4	0.966	1.15	3.43	2-4	0.968	1.13	3.41
3-4	0.962	1.05	2.96	3-4	0.966	1.00	2.91	3-4	0.967	0.99	2.90
4-3 (C_2)	0.991	0.85	2.72	4-3 (C_2)	0.991	0.85	2.72	4-3 (C_2)	0.991	0.85	2.72

[a] Radius in angstroms and $-\nabla^2\rho$ and ρ in atomic units.

Table 5. Ellipticity (ϵ), ρ, and $-\nabla^2\rho$ of the Carbon−Halogen (Fluorine, Chlorine, and Bromine) Bonds for Series 2 and Series 3[a]

	fluorine				chlorine				bromine			
molecule	bond	$\epsilon \times 10^2$	$\rho \times 10^1$	$-\nabla^2\rho \times 10^1$	bond	$\epsilon \times 10^2$	$\rho \times 10^1$	$-\nabla^2\rho \times 10^1$	bond	$\epsilon \times 10^2$	$\rho \times 10^1$	$-\nabla^2\rho \times 10^1$
2-1	C−F	7.36	2.49	1.44	C−Cl	5.09	1.94	2.84	C−Br	6.22	1.57	1.47
3-1	C−F	3.25	2.23	0.37	C−Cl	1.40	1.67	1.87	C−Br	1.26	1.36	1.05
2-2	C−F	6.58	2.48	1.21	C−Cl	5.61	1.91	2.72	C−Br	6.37	1.55	1.44
3-2	C−F	1.02	2.18	0.63	C−Cl	0.86	1.61	1.64	C−Br	0.85	1.31	0.92
2-3	C−F	5.66	2.47	1.31	C−Cl	5.73	1.90	2.64	C−Br	6.38	1.53	1.39
3-3	C−F	0.46	2.16	0.64	C−Cl	0.90	1.60	1.58	C−Br	0.76	1.28	0.84
2-4	C−F	6.12	2.48	1.20	C−Cl	5.71	1.91	2.72	C−Br	6.47	1.55	1.45
3-4	C−F	1.17	2.18	0.65	C−Cl	0.90	1.61	1.63	C−Br	0.86	1.31	0.91

[a] ϵ, ρ, and $-\nabla^2\rho$ in atomic units.

found between unsaturated aliphatic and aromatic compounds in terms of $-\nabla^2\rho$ and ρ values. For example, ethane (1-1) and chloroethene (2-1) exhibit larger radii and greater values of $-\nabla^2\rho$ and ρ than benzene (1-2) and chlorobenzene (2-2), respectively.

4.4. Characterization of the Carbon−Halogen and Carbon−Carbon Bonds. It is well-established that the electronic properties at the bond critical point provide extensive information about a chemical bond.[2] Therefore, we carried out the characterization of the bond critical points at the carbon−halogen bond in series 2 and series 3 for the three halogens in terms of ellipticity, ρ, and $-\nabla^2\rho$. The bond critical point at the carbon−carbon bond was also characterized, and a comparison with the carbon−carbon bonds in series 1 and series 4 was carried out. The data in Table 5 indicate that the ellipticities of carbon−halogen bonds are greater in systems where the resonance effect exists. For example, the ellipticity of the carbon−chlorine bond in chloroethene (2-1) is 5.09×10^{-2} au (atomic units), which is greater than the 1.40×10^{-2} au exhibited in chloroethane (3-1). Moreover, the ellipticities are greater in halogen unsaturated aliphatic systems (2-1) than in halogen aromatic systems (2-2, 2-3, and 2-4) for chlorine and bromine. The opposite behavior is observed for fluorine.

The electron density at the bond critical point is also greater in systems where the halogen resonance effect is present (series 2). Fluorine produces the greatest value of ρ at the bond critical point. There are no appreciable differences between halogen unsaturated aliphatic and halogen aromatic systems in terms of ρ at the carbon−halogen bond critical point. Similar behavior is observed for $-\nabla^2\rho$ except that chlorine produces the greatest values. $-\nabla^2\rho$ values for halogen unsaturated aliphatic systems are slightly greater than those of halogen aromatic systems for the three halogens.

Table 6 shows that series 2 exhibits greater values of ellipticities in the carbon−carbon double bond than series 1. For example, the ellipticity at the carbon−carbon double bond in fluoroethene (2-1) is 4.20×10^{-1} au, which is greater than the 3.32×10^{-1} au exhibited in ethene (1-1). This observation can be explained by the delocalization of one nonbonded maximum charge concentration by the resonance effect through the carbon−halogen bond, which contributes to an increase in the electron density in the π plane of the whole system.

ρ and $-\nabla^2\rho$ at the bond critical point of the carbon−carbon double bond are slightly greater in series 2 than in series 1, with fluorine yielding the greatest differences. Moreover, the carbon−carbon bond in series 2 exhibits considerably greater

Table 6. Ellipticity (ϵ), ρ, and $-\nabla^2\rho$ of the Carbon−Carbon Bond Adjacent to the Halogen for All Four Series[a]

fluorine				chlorine				bromine			
molecule	*e*	$\rho \times 10^1$	$-\nabla^2\rho \times 10^1$	molecule	*e*	$\rho \times 10^1$	$-\nabla^2\rho \times 10^1$	molecule	*e*	$\rho \times 10^1$	$-\nabla^2\rho \times 10^1$
1-1	3.3210^{-1}	3.44	10.27	1-1	3.3210^{-1}	3.44	10.27	1-1	3.3210^{-1}	3.44	10.27
2-1	4.2010^{-1}	3.52	10.72	2-1	3.7910^{-1}	3.47	10.37	2-1	3.6810^{-1}	3.47	10.34
3-1	4.1510^{-2}	2.54	6.13	3-1	1.6710^{-2}	2.48	5.80	3-1	9.3710^{-3}	2.48	5.78
4-1	2.2410^{-5}	2.41	5.49	4-1	2.2410^{-5}	2.41	5.49	4-1	2.2410^{-5}	2.41	5.49
1-2	2.0010^{-1}	3.08	8.59	1-2	2.0010^{-1}	3.08	8.59	1-2	2.0010^{-1}	3.08	8.59
2-2	2.6110^{-1}	3.17	9.10	2-2	2.3210^{-1}	3.10	8.65	2-2	2.2410^{-1}	3.10	8.60
3-2	4.3810^{-2}	2.53	6.03	3-2	2.0710^{-2}	2.47	5.70	3-2	1.4010^{-2}	2.47	5.67
4-2	6.6110^{-3}	2.39	5.31	4-2	6.6110^{-3}	2.39	5.31	4-2	6.6110^{-3}	2.39	5.31
1-3	2.4210^{-1}	3.20	9.13	1-3	2.4210^{-1}	3.20	9.13	1-3	2.4210^{-1}	3.20	9.13
2-3	3.0910^{-1}	3.28	9.58	2-3	2.7710^{-1}	3.21	9.14	2-3	2.6910^{-1}	3.20	9.08
3-3	4.4510^{-2}	2.53	6.04	3-3	2.4410^{-2}	2.48	5.73	3-3	1.5210^{-2}	2.46	5.65
4-3	5.9710^{-3}	2.40	5.34	4-3	5.9710^{-3}	2.40	5.34	4-3	5.9710^{-3}	2.40	5.34
1-3	2.4210^{-1}	3.20	9.13	1-3	2.4210^{-1}	3.20	9.13	1-3	2.4210^{-1}	3.20	9.13
2-4	3.1310^{-1}	3.28	9.59	2-4	2.8110^{-1}	3.22	9.17	2-4	2.7210^{-1}	3.21	9.12
3-4	4.3010^{-2}	2.53	6.04	3-4	1.9910^{-2}	2.48	5.72	3-4	1.3110^{-2}	2.47	5.69
4-3	5.9710^{-3}	2.40	5.34	4-3	5.9710^{-3}	2.40	5.34	4-3	5.9710^{-3}	2.40	5.34

[a] ϵ, ρ, and $-\nabla^2\rho$ in atomic units.

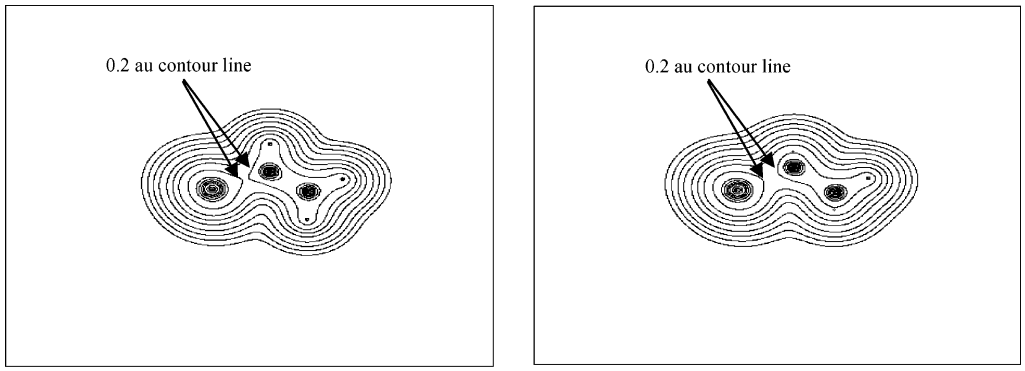


Figure 3. Contour map of the electron density in the plane that contains chlorine and the two carbons. (a) Chloroethene and (b) chloroethane.

values of ρ and $-\nabla^2\rho$ at the bond critical point than those in series 3. For example, fluoroethene (2-3) exhibits values of 3.52 and 10.72 au (ρ and $-\nabla^2\rho$), which are greater than the 2.54 and 6.13 au exhibited in fluoroethane, respectively. These facts can only be explained by the halogen resonance effect.

Furthermore, Figure 3 shows the contour map of the electron density in the plane that contains the halogen and the two carbons in chloroethane and chloroethene. Greater distortion of the 0.2 au contour line in chloroethene than that of chloroethane is clearly evident. Therefore, the halogen resonance effect produces a greater amount of charge in the area between the two atoms, which further demonstrates the donor character of the halogen resonance effect.

The ring critical points were also analyzed for the cyclic systems. The three halogens do not produce any appreciable effects on ρ and $-\nabla^2\rho$ at the ring critical points. Their values are very similar to those of the nonhalogen systems. Therefore, the presence of the halogens in the systems does not make any appreciable difference in the characteristics of the ring critical points whether the resonance effect exists or not.

4.5. Population Analysis. Table 7 lists the populations of the halogens and their adjacent carbons in series 2 and series 3. Clearly, the populations of chlorine and bromine in series 2 are smaller than those in series 3, whereas the population of fluorine remains constant at 9.62 au. Notice that the populations of chlorine and bromine are still greater than their respective atomic numbers (17 and 35). This suggests that the electron-withdrawing inductive effect of the halogens is stronger than the donor resonance effect. However, the fluorine systems do not follow the same behavior.

The populations of the carbons adjacent to the halogens are always greater in series 2 than in series 3. This is an expected result and restates the donor character of the halogen resonance effect. The resonance effect produced by the delocalization of one nonbonded maximum charge concentration of the VSCC of the halogens in series 2 donates charge to the adjacent carbon, increasing its population. It makes the populations of the carbons adjacent to the halogen in series 2 greater than that of the respective carbons in series 3 where only the electron-withdrawing inductive is present.

Table 7. Populations of the Halogens and Their Adjacent Carbons in Series 2 and Series 3[a]

		Population		
molecule	atoms	fluorine	chlorine	bromine
2-1	halogen	9.62	17.21	35.07
	carbon	5.56	5.97	6.11
3-1	halogen	9.62	17.28	35.18
	carbon	5.49	5.88	5.99
2-2	halogen	9.62	17.22	35.08
	carbon	5.55	5.96	6.10
3-2	halogen	9.62	17.30	35.20
	carbon	5.52	5.89	5.994
2-3	halogen	9.62	17.22	35.08
	carbon	5.55	5.96	6.10
3-3	halogen	9.62	17.30	35.21
	carbon	5.46	5.96	6.00
2-4	halogen	9.62	17.22	35.08
	carbon	5.55	5.96	6.10
3-4	halogen	9.62	17.30	35.20
	carbon	5.519	5.89	5.99

[a] Populations in atomic units.

Table 8. Delocalization Indexes for the Bonded Halogen−Carbon [δ(X,C)] in Series 2 and Series 3[a]

molecule	fluorine	chlorine	bromine
2-1	0.850	1.125	1.158
3-1	0.813	1.028	1.051
2-2	0.826	1.092	1.121
3-2	0.787	0.985	1.003
2-3	0.823	1.087	1.114
3-3	0.784	0.980	0.989
2-4	0.825	1.092	1.122
3-4	0.786	0.984	1.001

[a] Delocalization indexes in atomic units.

4.6. Delocalization Indexes. As explained above, δ(X,C) measures the sharing of electrons between a halogen and carbon. δ(X,C) is clearly greater in series 2 than in series 3 for chlorine and bromine (Table 8). The same behavior is observed in the case of fluorine, although the difference is not as great. These facts support the donor character of the halogen resonance effect in series 2, which produces a higher sharing of electrons between the three halogens and their bonded carbon than in series 3. On the other hand, fluorine and bromine exhibit the lowest and largest values of δ(X,C), respectively. In fact, the sharing of electrons between halogens and their bonded carbons in our systems is inversely related to the eletronegativity difference between the halogen and its bonded carbon. For example, δ(Br,C) is greater than δ(F,C) and δ(Cl,C) in both series, and the electronegativity difference increases from Br−C to F−C.

4.7. Electrophilic Aromatic Substitution in the α- and β-Halonaphthalenes. In their study of electrophilic aromatic substitution, Bader and Chang[21] showed that the Laplacian values at the saddle points, which link the carbon−carbon bonded charge concentration in substituted benzenes, predict the observed directing and activating−deactivating effects. These so-called link points exhibit the second-highest charge concentration of the VSCC of the atoms (second to the maximum charge concentration). In benzene, these points

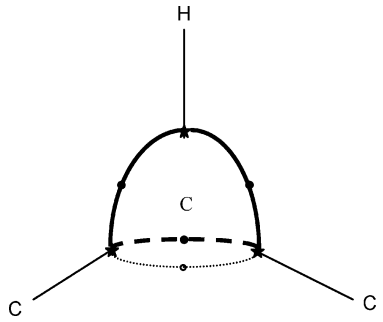


Figure 4. Atomic graphs describing the VSCC of a carbon in benzene. The maximum charge concentrations and link points are denoted by stars and dots, respectively. The solid link line is in the plane, the dashed link line is above the plane, and the gray link line is below the plane.

are above and below the plane of the ring. Therefore, it is reasonable to think of them as possible sites for electrophilic attack. The location of these link points in benzene is illustrated in Figure 4. In this paper, we have carried out a similar study of electrophilic aromatic substitution in the α- and β-halonaphthalenes.

As in the benzene case, there are two saddle points that link the two carbon−carbon bonded charge concentrations in the halonaphthalenes (one above and another below the plane of the ring). However, the saddle points which link the hydrogen−carbon maximum bonded charge concentrations are not always in the plane of the ring as in benzene (Figure 4). All carbons of the ring except the bridging carbons in the halonaphthalenes exhibit one hydrogen−carbon link point in the plane of the ring. Also, they exhibit two hydrogen−carbon link points out of the plane of the ring, one above and one below. It can be understood that one of the hydrogen−carbon link points that is in the plane in benzene is split into two hydrogen−carbon link points in the halonaphthalenes because of the lower symmetry. The location of the split of the link points alternates along the carbon chain. For example, there are four hydrogen−carbon link points in the region between C_1 and C_2 (the split of hydrogen−carbon link points for each carbon), and it repeats in the regions between C_3 and C_4, C_6 and C_7, and C_8 and C_9. On the other hand, there are two hydrogen−carbon link points in the region between C_2 and C_3 (the hydrogen−carbon link point in the plane for each carbon), and it repeats in the region between C_7 and C_8. Figure 5 illustrates the location of the link points in the α- and β-halonaphthalenes, and Table 9 provides $-\nabla^2\rho$ values of the carbon−carbon link points for the α- and β-halonaphthalenes. (The analysis of the hydrogen−carbon link points is not reported because it does not provide further insight. Their properties do not change appreciably throughout the ring, and also, they are generally similar to those of naphthalene for each halogen. Furthermore, their $-\nabla^2\rho$ values are noticeably lower than those for carbon−carbon link points. Thus, we focus on the carbon−carbon link points because they should play a major role as sites for electrophilic attack. We performed our analysis in terms of $-\nabla^2\rho$ to be consistent with the common practice in the literature, but it must be noted that the opposite convention was used by Bader and Chang.[21])

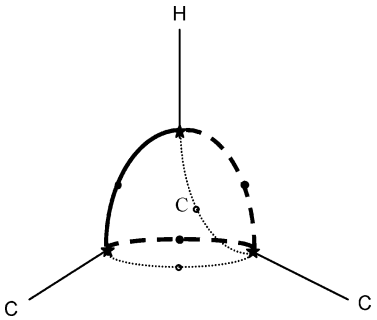


Figure 5. Atomic graphs describing the VSCC of a carbon in the halonaphthalenes. The maximum charge concentrations and link points are denoted by stars and dots, respectively. The solid link line is in the plane, the dashed link line is above the plane, and the gray link line is below the plane.

Table 9. Values of $-\nabla^2\rho$ for Carbon−Carbon Link Points in α- and β-Halonaphthalene[a]

α-halogen	F	Cl	Br	β-halogen	F	Cl	Br
C_2	0.159	0.150	0.149	C_1	0.165	0.156	0.154
C_3	0.135	0.136	0.136	C_3	0.150	0.146	0.146
C_4	0.145	0.139	0.136	C_4	0.136	0.136	0.137

[a] $-\nabla^2\rho$ in atomic units.

Experimental studies of electrophilic aromatic substitution[30] indicate that halogens at the α position (C_1) in naphthalene are C_2- and C_4-directing, whereas halogens at the β position (C_2) are C_1- and C_3-directing. As can be seen in the table, the link points with the largest values of $-\nabla^2\rho$ are in the VSCCs of C_2 and C_4 in the case of the α-halonaphthalenes. α-fluoronaphthalene exhibits the greatest values, 0.159 for C_2 and 0.145 for C_4, whereas α-bromonaphthalene exhibits the lowest values, 0.149 for C_2 and 0.136 for C_4. Thus, the electrophilic attack will preferentially occur at C_2 and C_4, following the trend F > Cl > Br as demonstrated by experiments. However, our results are not absolutely consistent with experimental results[30] because they predict that the electrophilic attack will be more preferentially directed to position C_2 than to position C_4 in the α-halonaphthalenes. Also, they predict a similar tendency of electrophilic aromatic substitution at the C_3 and C_4 positions except for the case of fluorine, where there is a noticeable preference of C_4 over C_3. The analysis for β-halonaphthalenes indicates a preference for C_1 over C_3 with the same trend as that observed in the α-halonaphthalenes (F > Cl > Br). These results are consistent with experimental results. Furthermore, the results predict faster electrophilic aromatic substitution at the C_1 position of the β-halonaphthalenes than at the C_2 position of the α-halonaphthalenes. These results are also consistent with experimental results.[30]

5. Conclusions

The VSCCs of the halogens in compounds where the halogens are bonded to a carbon−carbon single bond exhibit three nonbonded maximum charge concentrations (series 3). The location of these maxima can be considered to be tetrahedral. However, the VSCCs of the halogens bonded to a carbon−carbon double bond exhibit two nonbonded maximum charge concentrations in the sp^2 plane of the carbons (series 2). This suggests an overlapping or delocalization of one of the nonbonded maximum charge concentrations of the halogen into the π cloud of the carbon−carbon double bond by a resonance effect. The systems with the halogen resonance effect exhibit smaller radii and larger values of $-\nabla^2\rho$ and ρ at the maximum charge concentrations of the halogens than those in the systems where only the electron-withdrawing inductive effect of the halogen is acting.

In the case of the sp^2 carbons bonded to the halogen in series 2, their maximum charge concentrations bonded to the halogen exhibit smaller radii and greater values of $-\nabla^2\rho$ and ρ than those of the sp^3 carbons bonded to the halogen in series 3. Moreover, the maximum charge concentrations of the sp^2 carbons bonded to the other carbon that forms the carbon−carbon double bond in series 2 exhibit similar radii (in general, slightly smaller) and greater values of $-\nabla^2\rho$ and ρ than those of the sp^3 carbons bonded to the halogen in series 3. The values of $-\nabla^2\rho$ and ρ for these carbon−carbon-bonded maximum charge concentrations in the four series of molecules follow the trend 2 > 1 > 3 > 4.

The ellipticities, $-\nabla^2\rho$, and ρ at the bond critical points for the halogen−carbon bonds are greater in series 2 than in series 3. Fluorine and chlorine produce the largest values of ρ and $-\nabla^2\rho$, respectively. These facts clearly show the delocalization of charge from the halogens to the sp^2 carbon. In the case of the carbon−carbon bond, the ellipticities at the bond critical point are greater in series 2 than in series 1 (also true for series 3 and series 4). These results also suggest an overlapping or delocalization of one of the nonbonded maximum charge concentrations of the VSCC of the halogens into the π cloud of the carbon−carbon double bond by a resonance effect. The ρ and $-\nabla^2\rho$ values exhibit the same behavior.

The populations of chlorine and bromine in series 2 are smaller than in series 3 (larger than 17 and 35, respectively). This suggests that the electron-withdrawing inductive effect of the halogens is larger than the donor resonance effect. However, the fluorine systems do not follow the same behavior. In the case of the carbon bonded to the halogens, the populations are always greater in series 2 than in series 3, which is consistent with the donor character of the halogen resonance effect. Furthermore, the delocalization indexes between the halogen and its bonded carbon are larger in series 2 than in series 3, which shows that the halogen resonance effect contributes to the sharing of electrons between the halogens and their bonded carbon.

The locations of the carbon−carbon link points of the VSCC in α- and β-naphthalene is slightly different than in benzene. The analysis of these link points in terms of $-\nabla^2\rho$ shows that electrophilic aromatic substitution is more favorable in α-fluoronaphthalene than in α-chloronaphthalene and α-bromonaphthalene. The same conclusion holds for the β-halonaphthalenes. Also, the results indicate a preference for C_1 over C_3 in the β-halonaphthalenes. All these results are consistent with experimental results.

In summary, we report several observations that are consistent with the presence of the halogen resonance effect in compounds where the halogen is bonded to a carbon−

carbon double bond. These observations include the missing nonbonded maximum charge concentration in the VSCC of the halogens, the increase of $-\nabla^2\rho$ and ρ in the bonded maximum charge concentrations in the VSCC of the halogens and in the VSCC of their respective bonded carbons, the electronic properties at the BCPs, the atomic populations, and the delocalization indexes. Furthermore, our observations are consistent with experimental results for electrophilic aromatic substitution in halonaphthalenes.

Acknowledgment. The authors wish to acknowledge the financial support of the Natural Sciences and Engineering Research Council of Canada.

References

(1) Dewar, M. J. S. *Mol. Struct. Energ.* **1988**, *6*, 1−61.

(2) Bader, R. F. W. *Atoms in Molecules—A Quantum Theory*; Oxford University Press: New York, 1990.

(3) Bader, R. F. W. *Chem. Rev.* **1991**, *91*, 893−928.

(4) Bader, R. F. W. *Can. J. Chem.* **1998**, *76*, 973−988.

(5) Bader, R. F. W.; Nguyen-Dang, T. *Adv. Quantum Chem.* **1981**, *14*, 63−123.

(6) Popelier, P. L. A. Atoms in Molecules. An Introduction; Prentice Hall: London, 2000.

(7) Matta, C. F.; Gillespie, R. J. *J. Chem. Educ.* **2002**, *79*, 1141−1151.

(8) Bader, R. F. W.; Stephens, M. E. *J. Am. Chem. Soc.* **1975**, *97*, 7391−7399.

(9) Fradera, X.; Austen, M. A.; Bader, R. F. W. *J. Phys. Chem. A* **1999**, *103*, 304−314.

(10) Austen, M. A. A New Procedure for Determining Bond Orders in Polar Molecules, with Applications to Phosphorus and Nitrogen Containing Systems. Ph.D. Thesis, McMaster University, Hamilton, Canada, 2003.

(11) Bader, R. F. W. *Inorg. Chem.* **2001**, *40*, 5603−5611.

(12) Matta, C. F.; Hernandez-Trujillo, J. *J. Phys. Chem. A* **2003**, *107*, 7496−7504.

(13) Bader, R. F. W.; Streitwieser, A.; Neuhaus, A.; Laidig, K. E.; Speers, P. *J. Am. Chem. Soc.* **1996**, *118*, 4959−4965.

(14) González, M. J.; Mosquera, R. *J. Phys. Chem. A* **2003**, *107*, 5361−5367.

(15) Okulik, N.; Peruchena, N. M.; Esteves, P. M.; Mota, C. J. A.; Jubert, A. *J. Phys. Chem. A* **1999**, *103* (42), 8491−8495.

(16) Okulik, N. B.; Sosa, L. G.; Esteves, P. M.; Mota, C. J. A.; Jubert, A. H.; Peruchena, N. M. *J. Phys. Chem. A* **2002**, *106* (8), 1584−1595.

(17) Grabowski, S. J. *J. Mol. Struct.* **2001**, *562* (1−3), 137−143.

(18) Grabowski, S. J. *J. Phys. Org. Chem.* **2003**, *16* (10), 797−802.

(19) Gilli, P.; Bertolasi, V.; Pretto, L.; Lycka, A.; Gilli, G. *J. Am. Chem. Soc.* **2002**, *124* (45), 13554−13567.

(20) Borbulevych, O. Y.; Clark, R. D.; Romero, A.; Tan, L.; Antipin, M. Y.; Nesterov, V. N.; Cardelino, B. H.; Moore, C. E.; Sanghadasa, M.; Timofeeva, T. V. *J. Mol. Struct.* **2002**, *604* (1), 73−86.

(21) Bader, R. F. W.; Chang, C. *J. Phys. Chem. A* **1989**, *93*, 2946−2956.

(22) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. J. A.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; 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. W.; Gill, P. M.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03*, revision B.03.; Gaussian Inc.: Pittsburgh, PA.

(23) Bader, R. F. W. *AIMPAC*. http://www.chemistry.mcmaster.ca/aimpac/.

(24) Biegler-König, F. W.; Bader, R. F. W.; Tang, T.-H. *J. Comput. Chem.* **1982**, *13*, 317−328.

(25) Matta, C. F. *AIMDELOC: Program to calculate AIM localization and delocalization indexes (QCPE0802)*; Quantum Chemistry Program Exchange: Indiana University, IN. http://qcpe.chem.indiana.edu/.

(26) Bader, R. F. W.; Gillespie, R. J.; MacDougall, P. J. *J. Am. Chem. Soc.* **1988**, *110*, 7329−7336.

(27) Bader, R. F. W.; Heard, G. L. *J. Chem. Phys.* **1999**, *111*, 8789−8798.

(28) Castillo, N.; Boyd, R. *J. Chem. Phys. Lett.* **2005**, *403*, 47−54.

(29) Popelier, P. L. A.; Burke, J.; Malcolm, N. O. J. *Int. J. Quantum Chem.* **2003**, *92*, 326−336.

(30) Taylor, R. *Electrophilic Aromatic Substitution*; John Wiley & Sons Ltd: England, 1990.

CT050238J

JCTC Journal of Chemical Theory and Computation

Proton Affinities of Anionic Bases: Trends Across the Periodic Table, Structural Effects, and DFT Validation

Marcel Swart and F. Matthias Bickelhaupt*

Theoretische Chemie, Scheikundig Laboratorium der Vrije Universiteit, De Boelelaan 1083, NL-1081 HV Amsterdam, The Netherlands

Received October 3, 2005

Abstract: We have carried out an extensive exploration of the gas-phase basicity of archetypal anionic bases across the periodic system using the generalized gradient approximation of density functional theory (DFT) at BP86/QZ4P//BP86/TZ2P. First, we validate DFT as a reliable tool for computing proton affinities and related thermochemical quantities: BP86/QZ4P//BP86/TZ2P is shown to yield a mean absolute deviation of 1.6 kcal/mol for the proton affinity at 0 K with respect to high-level ab initio benchmark data. The main purpose of this work is to provide the proton affinities (and corresponding entropies) at 298 K of the anionic conjugate bases of all main-group-element hydrides of groups 14−17 and periods 2−6. We have also studied the effect of stepwise methylation of the protophilic center of the second- and third-period bases.

1. Introduction

Designing new (and optimizing existing) approaches and routes in organic synthesis requires knowledge of the thermochemistry involved in the targeted reactions. In this context, the proton affinity (PA) of a reactant or intermediate species B^- often plays an important role. This thermochemical quantity is defined as the enthalpy change associated with dissociation of the conjugate acid (eq 1):[1−5]

$$BH \rightarrow B^- + H^+ : \Delta H = PA \qquad (1)$$

Overall reaction enthalpies and reaction barriers (and thus reaction rates) are related to the PA, as soon as proton transfer occurs somewhere along the cascade of elementary steps of a reaction mechanism. This is often the case, as proton transfer is ubiquitous in organic reaction mechanisms, either as simple proton transfer (PT) or as part of a more complex chemical transformation, for example, base-induced elimination reactions that may compete with nucleophilic substitution.[6]

Here, we focus on the proton affinities of anionic bases in the gas phase. Gas-phase proton affinities are obviously directly applicable to gas-phase chemistry,[3,4] but they are also relevant for chemistry occurring in the condensed phase.[1,7] On one hand, they reveal the intrinsic basicity of the protophilic species involved, and thus, they shed light on how this property is affected by the solvent. On the other hand, they can serve as a universal, solvent-independent framework of reference, from which the actual basicity of a species in solution can be obtained through an (empirical) correction for the particular solvent under consideration.[8] Experimental gas-phase proton affinities are well-known for neutral and cationic organic bases.[3,4] Less information is available for anionic bases, in particular, for anionic bases with a heavier, that is, third- and higher-period, atom as the protophilic center.[4,5]

The present study has three purposes. First, we wish to evaluate the performance of various popular exchange-correlation functionals of density functional theory (DFT)[9] ranging from the local density approximation (LDA) via the generalized gradient approximation (GGA) and hybrid functionals to meta-GGA functionals.[9,10] This is done by computing the 0 K reaction enthalpies $\Delta_{acid}H_0$ of reaction 1 (i.e., PA^0 values) for a test set of 17 bases for which highly accurate benchmark data are available.[11,12] It is anticipated here that the well-known BP86[13] functional performs very reasonably, in fact, even slightly better than the B3LYP[14] hybrid functional. Second, we aim at setting up a complete description of the proton affinities of the conjugate bases XH_n^- of all archetypal main-group-element hydrides XH_{n+1}

* Corresponding author fax: +31-20-59 87617; e-mail: fm.bickelhaupt@few.vu.nl.

10.1021/ct0502460 CCC: $33.50
Published on Web 12/24/2005

of groups 14−17 and periods 2−6. Third, we have studied the influence of stepwise methylation of the protophilic center X in species $(CH_3)_mXH_{n-m}^-$ (for periods 2 and 3), for example, SiH_3^-, $CH_3SiH_2^-$, $(CH_3)_2SiH^-$, and $(CH_3)_3Si^-$. In addition to our computed BP86/QZ4P//BP86/TZ2P values for the 298 K reaction enthalpy $\Delta_{acid}H_{298}$ of all acids BH in eq 1 (that is, the proton affinity PA of all bases B^-), we also report the corresponding 298 K reaction entropies ($\Delta_{acid}S_{298}$, provided as $-T\Delta_{acid}S_{298}$ values) and 298 K reaction free energies ($\Delta_{acid}G_{298}$).

Note that the series of, in total, 41 bases investigated in this study covers large parts of the periodic system as well as a number of important structural themes occurring in organic chemistry. To the best of our knowledge, this series of bases has never before been studied, in its full range, consistently with one and the same method, either experimentally or theoretically. We anticipate that the very consistency of our approach makes our data particularly suitable for inferring accurate *trends* in thermochemistry across the periodic system.

2. Methods

All calculations were performed with the Amsterdam Density Functional program developed by Baerends and others.[15,16] Molecular orbitals were expanded using two different large, uncontracted sets of Slater-type orbitals: TZ2P and QZ4P.[17] The TZ2P basis is of triple-ζ quality, augmented by two sets of polarization functions (d and f on heavy atoms; 2p and 3d on H). The QZ4P basis, which contains additional diffuse functions, is of quadruple-ζ quality, augmented by four sets of polarization functions (two d and f on heavy atoms; two 2p and two 3d sets on H). Core electrons (e.g., 1s for second period, 1s2s2p for third period, 1s2s2p3s3p for fourth period, 1s2s2p3s3p3d4s4p for fifth period, and 1s2s2p3s3p3d4s4p4d for sixth period) were treated by the frozen core approximation.[16] An auxiliary set of s, p, d, f, and g Slater-type orbitals was used to fit the molecular density and to represent the Coulomb and exchange potentials accurately in each self-consistent field (SCF) cycle. Scalar relativistic corrections were included self-consistently using the zeroth order regular approximation.[18]

Energies and gradients were calculated using LDA (Slater exchange and VWN[19] correlation) with nonlocal corrections[13] due to Becke (exchange) and Perdew (correlation) added self-consistently. This is the BP86 density functional, which is one of the three best DFT functionals for the accuracy of geometries,[20] with an estimated unsigned error of 0.009 Å in combination with the TZ2P basis set. The restricted and unrestricted formalisms were used for closed-shell and open-shell species, respectively.

The energies of a range of other popular DFT functionals were calculated in a post-SCF fashion using the BP86/QZ4P//BP86/TZ2P orbitals and densities, to estimate the influence of the choice of DFT functional. These functionals include LDA,[19] GGAs (BLYP, PBE, OLYP, and HCTH/407),[13,21,22] meta-GGAs (VS98, BLAP3, TPSS, and τ-HCTH),[13,23,24] and hybrid functionals (TPSSh, O3LYP, PBE0, B97, B1PW91, and τ-HCTHh).[21,24,25] The resulting proton affinities at 0 K and corresponding deviations from high-level theory data are provided in Table S1 in the Supporting Information.

Geometries were optimized using analytical gradient techniques until the maximum gradient component was less than 1.0×10^{-4} atomic units (see Table S2 in the Supporting Information). Vibrational frequencies were obtained through numerical differentiation of the analytical gradients.[16] The enthalpy correction to the electronic energy of the systems was calculated from the vibrational frequencies using standard thermochemistry relations;[26] for example, enthalpy corrections at 298.15 K and 1 atm (ΔH_{298}) were calculated according to

$$\Delta H_{298} = \Delta E_{trans,298} + \Delta E_{rot,298} + \Delta E_{vib,0} + \Delta(\Delta E_{vib,0})_{298} + \Delta(pV)$$

Here, $\Delta E_{trans,298}$, $\Delta E_{rot,298}$, and $\Delta E_{vib,0}$ are the differences between the reactant (i.e., BH, the conjugate acid) and products (i.e., $B^- + H^+$, the anionic base and the proton) in translational, rotational, and zero-point vibrational energy, respectively; $\Delta(\Delta E_{vib,0})_{298}$ is the change in the vibrational energy difference as one goes from 0 to 298.15 K. The vibrational energy corrections are based on our frequency calculations. The molar work term $\Delta(pV)$ is $(\Delta n)RT$; $\Delta n = +1$ for one reactant BH dissociating to two products (B^- and H^+). Thermal corrections for the electronic energy are neglected.

3. Results and Discussion

3.1. Benchmarking and Validation of DFT. We begin with an extensive exploration of the performance of various density functionals covering LDA, GGA, meta-GGA, and hybrid DFT. To this end, we have computed the proton affinity at 0 K ($PA^0 = \Delta_{acid}H_0$) for a series of 17 anionic bases B^-, shown in Table 1, for which extremely accurate experimental data (with uncertainties of only 0.003 up to 0.7 kcal/mol)[11,12] and ab initio theoretical benchmark values[12] are available. This series of bases covers PA^0 values ranging from 322.6 for Br^- through 415.2 kcal/mol for CH_3^- (see Table 1, exptl.). Table 1 compares our results for BP86, which emerges as one of the best functionals (vide infra), with earlier B3LYP/aug-cc-pVTZ and CCSD(T)/aug-cc-pVQZ//B3LYP/aug-cc-pVTZ benchmark calculations as well as with experimental results.[11,12] Our results for the other density functionals are collected in Table S1 of the Supporting Information. As can be seen in Table 1, the CCSD(T) benchmark and experimental PA^0 values agree excellently, the latter showing a mean absolute deviation (MAD) from the former of only 0.4 kcal/mol.

First, we have explored with BP86 the effect of carrying out the DFT calculations with the TZ2P versus the very large QZ4P basis set (see Section 2). Already, with the "smaller" TZ2P basis, that is, at BP86/TZ2P, we find PA^0 values in reasonable agreement with the CCSD(T) benchmark, showing a MAD value of 2.8 kcal/mol with respect to the latter (see Table 1). There is one qualitative disagreement between BP86/TZ2P, on one hand, and CCSD(T) and the experimental results, on the other hand: the former yields the vinyl anion $C_2H_3^-$ as slightly less basic than the amide anion NH_2^-,

Table 1. Computed[a,b,c] and Experimental[c] Proton Affinities at 0 K PA^0 (in kcal/mol) of Anionic Bases

base	BP86/TZ2P[a]	BP86/QZ4P[b]	B3LYP[c]	CCSD(T)[c,d]	exptl.[c]
CH_3^-	416.9	412.5	414.2	415.3	415.2 ± 0.7
$C_2H_3^-$	405.7	403.8	405.9	406.7	407.4 ± 0.3
NH_2^-	408.1	402.1	401.2	402.2	401.9 ± 0.1
$C_6H_5^-$	398.5	397.2	399.9		399.6 ± 0.4
H^-	400.1	398.3	398.1	399.6	399.5 ± 0.003
HCO^-	394.9	388.9	390.4	393.2	393.1 ± 0.1
OH^-	395.0	389.0	387.3	389.4	389.1 ± 0.02
CH_3O^-	376.6	375.4	377.8	381.0	380.7 ± 0.6
$CH_3CH_2O^-$	373.0	372.1	375.0		377.6 ± 0.7
C_2H^-	375.4	375.3	376.3	376.4	376.9 ± 0.1
$(CH_3)_2CHO^-$	370.9	370.2	373.3		375.4 ± 0.6
$(CH_3)_3CO^-$	370.3	369.9	372.7		374.6 ± 0.5
F^-	375.3	371.8	367.1	370.9	370.4 ± 0.003
SH^-	351.2	350.2	349.0	350.1	350.1 ± 0.01
CN^-	354.8	349.4	348.8	349.3	349.5 ± 0.2
Cl^-	334.9	332.6	330.8	332.7	332.5 ± 0.002
Br^-	323.4	323.4	321.6	324.3	322.6 ± 0.05
MAD/MD wrt CCSD(T)[e]	*2.8/1.6*	*1.6/−1.4*	*1.7/−1.7*	*0*	*0.4/−0.2*
MAD/MD wrt exp.[f]	*2.3/−2.0*	*2.3/−2.0*	*1.6/−1.6*	*0.4/0.2*	*0*

[a] This work: BP86/TZ2P energies with ZPE correction at BP86/TZ2P. [b] This work: BP86/QZ4P//BP86/TZ2P energies with ZPE correction at BP86/TZ2P. [c] B3LYP/aug-cc-pVTZ from ref 12. [d] CCSD(T)/aug-cc-pVQZ//B3LYP/aug-cc-pVTZ from ref 12. [e] Mean absolute deviation (MAD)/mean deviation (MD) relative to CCSD(T). [f] Mean absolute deviation (MAD)/mean deviation (MD) relative to experiment.

whereas the two latter yield the reverse order, that is, NH_2^- as more basic than $C_2H_3^-$ (see Table 1).

Further improvements can be achieved by going from the TZ2P to the QZ4P basis set. The QZ4P basis set is not only more flexible and better polarized it also contains more diffuse functions. One may, therefore, expect an improved description of reaction 1 in which we go from a neutral species BH to two (oppositely) charged species $B^- + H^+$. In particular, the description of the expanding density ("breathing orbitals") at the protophilic center (e.g., nitrogen in NH_3/NH_2^-) benefits from going from the TZ2P to the QZ4P basis set. Thus, single-point calculations were done at BP86/QZ4P using the BP86/TZ2P geometries and enthalpy corrections. At this level of theory, that is, at BP86/QZ4P//BP86/TZ2P, we achieve a significant improvement of the MAD, which drops to 1.6 kcal/mol relative to the CCSD(T) benchmark data, which is comparable to the MAD value of 1.7 kcal/mol obtained earlier at B3LYP/aug-cc-pVTZ[12] (see Table 1). Furthermore, BP86/QZ4P//BP86/TZ2P yields correct relative PA^0 values over the entire range of bases. None of the other density functionals (see Section 2) performs better than this BP86 approach, with MAD values ranging from 1.6 kcal/mol for PBE (the only other functional achieving this small MAD value) via 2.6 and 3.0 kcal/mol for BLYP and OLYP through 7.8 and 8.1 kcal/mol for O3LYP and LDA (see Table S1 in the Supporting Information).

We have verified that neither the geometry nor the enthalpy corrections differ significantly if they are also computed with the QZ4P basis set (data not shown). Thus, the full BP86/QZ4P//BP86/QZ4P energies differ by merely 0.04 kcal/mol or less compared to the BP86/QZ4P//BP86/TZ2P energies (tested for NH_2^- and F^-), while the enthalpy corrections differ by only 0.05 kcal/mol or less (tested for CH_3O^- and F^-).

In conclusion, BP86/QZ4P//BP86/TZ2P (with BP86/TZ2P enthalpy corrections) emerges as a reliable approach for studying trends in the basicity of anionic bases in the following section.

3.2. Proton Affinities of Main-Group-Element Hydrides. Using the BP86/QZ4P//BP86/TZ2P approach (see previous section), we have computed the proton affinities at 298 K ($PA = \Delta_{acid}H_{298}$) and the corresponding entropies $\Delta_{acid}S_{298}$ (provided as $-T\Delta_{acid}S_{298}$ values) and reaction free energies $\Delta_{acid}G_{298}$ of the anionic conjugate bases of all main-group-element hydrides of groups 14−17 and periods 2−6. The results are summarized in Table 2 and Figure 1.

Along the series of second-period bases, we obtain the well-known trend of a decreasing basicity as the PA falls from 414 to 404 to 390 to 373 kcal/mol along CH_3^-, NH_2^-, OH^-, and F^- (see Table 2 and Figure 1). In each of the four groups (14−17), the PA decreases if one descends the periodic table. The largest reduction in PA occurs from the second to the third period. In group 14, for example, the PA decreases from 414 to 369 to 356 to 342 to 324 kcal/mol along CH_3^-, SiH_3^-, GeH_3^-, SnH_3^-, and PbH_3^- (see Table 2). Interestingly, the changes in PA descending group 14 are significantly larger than in the other groups, 15−17. Thus, as can be seen in Figure 1, the trend of a monotonic decrease in PA along the second (P2) and, already to a lesser extent, the third period (P3) changes for the fourth through sixth periods (P4, P5, and P6) into a trend where the PA along a period first *increases* from group 14 to 15 and then decreases again along groups 15, 16, and 17. The corresponding reaction entropies yield a relatively small (but not entirely constant) contribution $-T\,\Delta_{acid}S_{298}$ of −5 to −9 kcal/mol for 298 K. As a consequence, the Gibbs free energies $\Delta_{acid}G_{298}$ show the same trends as the corresponding PA values (see Table 2).

Table 2. Thermodynamic Acidity Properties (kcal/mol) for Anionic Bases $Me_mH_{n-m}X^-$ at $T = 298$ K[a]

group 14				group 15				group 16				group 17			
base	ΔH	$-T\Delta S$	ΔG	base	ΔH	$-T\Delta S$	ΔG	base	ΔH	$-T\Delta S$	ΔG	base	ΔH	$-T\Delta S$	ΔG
							Period 2								
CH_3^-	414.0	−8.3	405.7	NH_2^-	403.6	−7.5	396.0	OH^-	390.2	−6.6	383.6	F^-	372.7	−5.7	366.9
$MeCH_2^-$	415.6	−8.7	406.8	$MeNH^-$	398.3	−7.4	390.8	MeO^-	376.6	−6.5	370.1				
Me_2CH^-	410.8	−7.9	402.9	Me_2N^-	387.1	−6.9	380.2								
Me_3C^-	403.9	−7.3	396.6												
							Period 3								
SiH_3^-	369.1	−8.3	360.8	PH_2^-	366.2	−7.5	358.6	SH^-	351.4	−6.4	345.0	Cl^-	333.5	−5.4	328.2
$MeSiH_2^-$	376.2	−8.3	367.9	$MePH^-$	372.9	−7.6	365.4	MeS^-	357.5	−6.4	351.0				
Me_2SiH^-	381.2	−8.0	373.1	Me_2P^-	377.4	−7.2	370.2								
Me_3Si^-	383.7	−7.5	376.2												
							Period 4								
GeH_3^-	355.9	−8.3	347.6	AsH_2^-	358.1	−7.5	350.6	SeH^-	341.9	−6.3	335.6	Br^-	324.3	−5.2	319.1
							Period 5								
SnH_3^-	342.0	−8.2	333.9	SbH_2^-	348.5	−7.5	341.0	TeH^-	332.3	−6.2	326.1	$I-$	315.9	−5.1	310.8
							Period 6								
PbH_3^-	324.4	−8.1	316.3	BiH_2^-	345.0	−7.5	337.5	PoH^-	329.2	−6.2	323.0	At^-	313.5	−5.0	308.5

[a] Computed at BP86/QZ4P//BP86/TZ2P for the reaction $Me_mH_{n-m}XH \rightarrow H^+ + Me_mH_{n-m}X^-$ with $n = 3$, 2, 1, and 0 for groups 14, 15, 16, and 17, respectively.

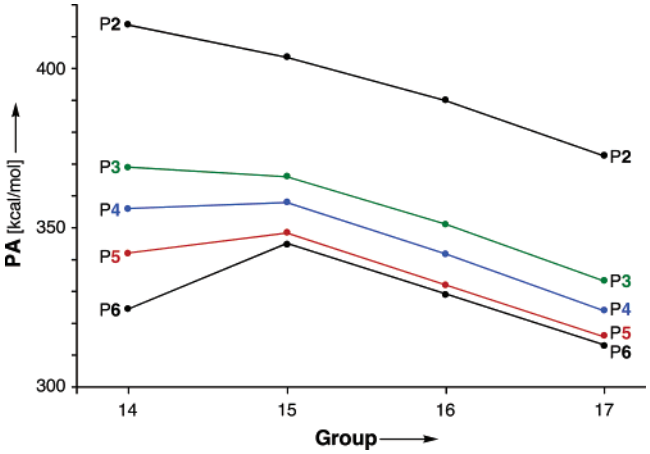
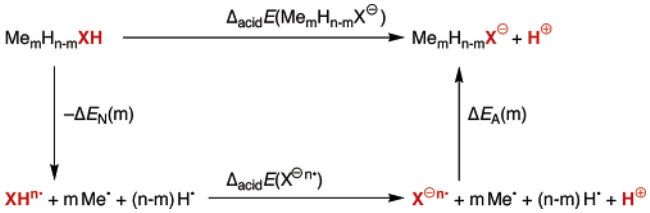


Figure 1. Proton affinities PA (at 298 K) of the anionic conjugate bases of main-group-element hydrides of groups 14−17 and periods 2−6 (P2−P6), computed at BP86/QZ4P//BP86/TZ2P.

3.3. Methyl Substituent Effects. Finally, we have studied the effect on the PA of a stepwise replacement of all hydrogen atoms in second- and third-period anionic bases H_nX^- by methyl substituents, that is, by a stepwise increase of m from 0 to n in $Me_mH_{n-m}X^-$ ($n = 3$, 2, 1, and 0 for group 14, 15, 16, and 17, respectively). The results are collected in Table 2. Strikingly, the PA of the second-period bases *decreases* while that of the third-period bases *increases* with the number of methyl substituents. For example, along NH_2^-, $MeNH^-$, and Me_2N^-, the PA decreases from 404 to 398 to 387 kcal/mol, which is in agreement with experimental gas-phase data,[4] whereas along PH_2^-, $MePH^-$, and Me_2P^-, it increases from 366 to 373 to 377 kcal/mol (see Table 2). This may seem to suggest that a methyl substituent stabilizes the second-period base (e.g., NH_2^-) compared to a hydrogen atom and that it destabilizes a third-row base (e.g., PH_2^-).

Scheme 1. Relationship between PA Values and Methyl Substituent Effect (see Table 3)



This is, however, incorrect, as follows from a more detailed analysis.

To trace the origin of the opposite methyl-substituent effects on second- versus third-row bases, we have decomposed the proton-affinity energy $\Delta_{acid}E$, associated with acid dissociation of the conjugate acid $Me_mH_{n-m}XH$, into three partial reactions, as shown in the thermochemical cycle of Scheme 1.

The first step, which is associated with an energy change $-\Delta E_N(m)$, is the dissociation of all substituents of the neutral, conjugate acid $Me_mH_{n-m}XH$ (but not the acidic proton) to form the n-fold radical $XH^{n\bullet}$ (see Scheme 1; m = number of methyl substituents). The energy ΔE_N is the stabilization of the protophilic center X in the neutral conjugate acid $Me_mH_{n-m}XH$ by all substituents, that is, the interaction with m methyl groups (Me) and n hydrogen atoms (H). Next, the unsubstituted acid $XH^{n\bullet}$ is dissociated into $X^{-n\bullet} + H^+$ (see Scheme 1). The corresponding reaction energy is the proton-affinity energy $\Delta_{acid}E$ of the anionic base $X^{-n\bullet}$. The third and last step, which is associated with an energy change $\Delta E_A(m)$, is the addition of all substituents to the protophilic center $X^{-n\bullet}$ to form the base $Me_mH_{n-m}X^-$ (see Scheme 1; m = number of methyl substituents). The energy ΔE_A is the stabilization of the protophilic center X in the anionic base $Me_mH_{n-m}X^-$ by all substituents, that is, the interaction with m methyl groups (Me) and $n - m$ hydrogen atoms (H). The

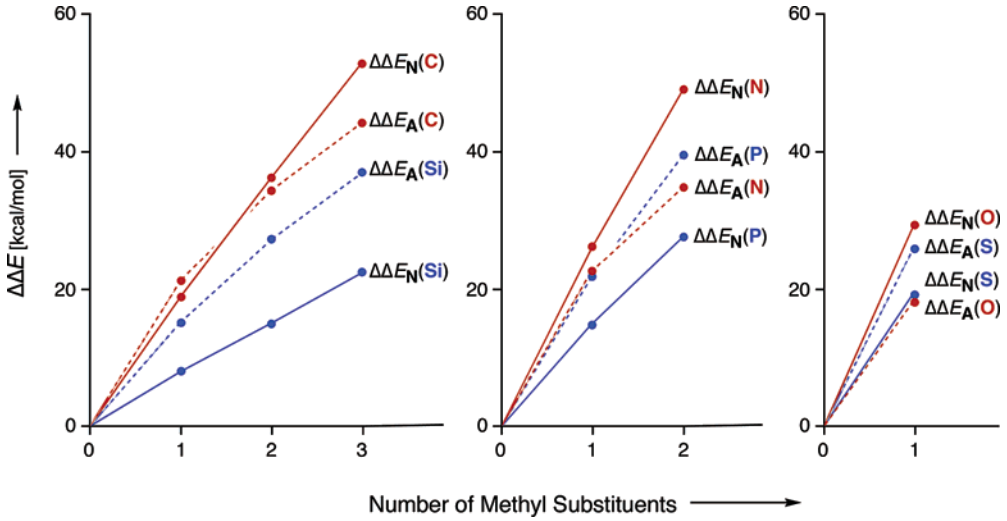


Figure 2. Effect of methyl substitution on the energy $\Delta E_N(X)$ of main-group-element hydrides H_nXH and the energy $\Delta E_A(X)$ of their anionic conjugate bases H_nX^- (see Scheme 1), computed at BP86/QZ4P//BP86/TZ2P (see Table 3).

Table 3. Analysis of the Methyl-Substituent Effect on PA Energies $\Delta_{acid}E$ in Terms of the Partial Reactions in Scheme 1[a]

base	ΔE_N	ΔE_A	$\Delta_{acid}E$	base	ΔE_N	ΔE_A	$\Delta_{acid}E$
$C^{\cdot\cdot\cdot-}$			350.00	$Si^{\cdot\cdot\cdot-}$			314.65
CH_3^-	−357.13	−284.65	422.48	SiH_3^-	−289.11	−228.89	374.87
$MeCH_2^-$	−338.30	−263.41	424.89	$MeSiH_2^-$	−281.28	−213.96	381.97
Me_2CH^-	−320.94	−250.34	420.60	Me_2SiH^-	−274.14	−201.77	387.02
Me_3C^-	−304.20	−240.36	413.84	Me_3Si^-	−266.91	−191.94	389.62
$N^{\cdot\cdot-}$			402.78	$N^{\cdot\cdot-}$			369.69
NH_2^-	−217.43	−208.39	411.82	PH_2^-	−169.66	−167.92	371.43
$MeNH^-$	−191.32	−185.99	408.11	$MePH^-$	−154.92	−146.19	378.42
Me_2N^-	−168.43	−173.59	397.62	Me_2P^-	−142.01	−128.49	383.21
$O^{\cdot-}$			387.48	$S^{\cdot-}$			353.36
OH^-	−126.55	−117.21	396.82	SH^-	−96.49	−94.08	355.77
MeO^-	−97.47	−99.11	385.84	MeS^-	−77.48	−68.27	362.57

[a] Computed at BP86/QZ4P//BP86/TZ2P. See also Figure 2.

computed proton-affinity energies' values $\Delta_{acid}E$, ΔE_N, and ΔE_A are collected in Table 3.

The relationship between the proton-affinity energy $\Delta_{acid}E(Me_mH_{n-m}X^-)$ of the anionic base and the other energy terms of the thermochemical cycle of Scheme 1 is summarized in eq 2 (m = number of methyl substituents):

$$\Delta_{acid}E(Me_mH_{n-m}X^-) = \Delta_{acid}E(X^{-n\bullet}) + \Delta E_A(m) - \Delta E_N(m) \quad (2)$$

Thus, the proton-affinity energy $\Delta_{acid}E(Me_mH_{n-m}X^-)$ of the base $Me_mH_{n-m}X^-$ is determined by the proton-affinity energy $\Delta_{acid}E(X^{-n\bullet})$ of the unsubstituted and deprotonated protophilic center $X^{-n\bullet}$ plus the *difference* in stabilization $\Delta E_A(m)$ of $X^{-n\bullet}$ in the base $Me_mH_{n-m}X^-$ and the stabilization $\Delta E_N(m)$ of $XH^{n\bullet}$ in $Me_mH_{n-m}XH$ by m methyl (Me) and n hydrogen substituents $(H)^{n\bullet}$. The methyl-substituent effect on the proton-affinity energy, that is, the change $\Delta\Delta_{acid}E(m)$ in this value if one goes from 0 to m methyl substituents, therefore, depends not only on the change $\Delta E_A(m) - \Delta E_A(0)$ in stabilization of the anionic base *but also on* the change $\Delta E_N(m) - \Delta E_N(0)$ *in stabilization of the neutral conjugate acid*.

This is the key to understanding the true origin of the opposite methyl-substituent effects on the PA of second- and third-period bases. In Figure 2, we have plotted the changes in stabilization by the substituents $\Delta\Delta E_A = \Delta E_A(m) - \Delta E_A(0)$ and $\Delta\Delta E_N = E_N(m) - \Delta E_N(0)$ for the second- and third-period bases $Me_mH_{n-m}X^-$ and their conjugate acids $Me_mH_{n-m}XH$. Now it is clear that introducing a methyl substituent leads consistently, in all cases, to a destabilization of the system. Thus, the reason second-period bases become less basic and third-period bases more basic is *not* that a methyl group stabilizes second-period bases and destabilizes the third-period bases. In fact, the introduction of methyl groups *destabilizes*, in all cases, both the base and the conjugate acid. The opposite trends in basicity originate from the fact that a *second-period* base is destabilized less than its corresponding conjugate acid, whereas a *third-period base is destabilized more* than its conjugate acid. This behavior is reminiscent of the situation of the halomethyl anions CH_2X^-, the PA of which decreases along X = F, Cl, Br, and I. Very recently,[27] we have shown that this is not because of increasing α-stabilization of CH_2X^- by X. Rather, the latter continuously *decreases* along the series, *but more* slowly so than the α-*stabilization of the conjugate acid* CH_3X.

4. Conclusions

BP86, B3LYP, and PBE emerge from our exploration of 41 model systems as sound and computationally efficient alternatives to highly correlated ab initio methods for computing proton affinities and related thermochemical quantities of anionic bases across the periodic table. The BP86/QZ4P//BP86/TZ2P approach achieves a mean absolute deviation of 1.6 kcal/mol for the proton affinity at 0 K ($\Delta_{acid}H_0$) with respect to high-level ab initio benchmark data. This is slightly more accurate than B3LYP, and in combination with its higher computational efficiency, this makes us recommend the above BP86 approach for obtaining accurate proton affinities of organic and inorganic species that either escape direct experimental observation or are computationally too demanding for highly correlated ab initio methods such as CCSD(T).

The proton affinity along the archetypal second-period bases CH_3^-, NH_2^-, OH^-, and F^- decreases as valence 2p atomic orbitals of the protophilic atom become more compact and stable. This well-known trend changes if one descends in the periodic system to higher periods. In each group, the proton affinity decreases, but it does so significantly more pronouncedly down group 14 than down the other groups. This causes the proton affinities along third- and higher-period bases to first increase from group 14 to group 15 and then to decrease again until group 17.

Introducing methyl substituents at the protophilic center has opposite effects on second-period and third-period anionic bases: while the former become less basic as the number of methyl substituents increases (e.g., along NH_2^-, $MeNH^-$, and Me_2N^-), the latter become more basic (e.g., along PH_2^-, $MePH^-$, and Me_2P^-). Interestingly, the reason for this is *not* that a methyl group stabilizes second-period bases and destabilizes the third-period bases. In fact, the introduction of methyl groups *destabilizes*, in all cases, both the base and the conjugate acid. The opposite trends in basicity originate from the fact that a *second-period base* is destabilized less than its corresponding conjugate acid, whereas a *third-period base is destabilized more* than its conjugate acid.

Acknowledgment. We thank the Netherlands organization for Scientific Research (NWO−CW) for financial support and Prof. N. M. M. Nibbering for helpful discussions.

Supporting Information Available: Proton affinities for 17 bases computed with various density functionals and Cartesian coordinates of all species occurring in this study. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Bon, R. S.; van Vliet, B.; Sprenkels, N. E.; Schmitz, R. F.; de Kanter, F. J. J.; Stevens, C. V.; Swart, M.; Bickelhaupt, F. M.; Groen, M. B.; Orru, R. V. A. *J. Org. Chem.* **2005**, *70*, 3542.

(2) Gal, J.-F.; Maria, P.-C.; Raczynska, E. D. *J. Mass Spectrom.* **2001**, *36*, 699.

(3) Hunter, E. P. L.; Lias, S. G. *J. Phys. Chem. Ref. Data* **1998**, *27*, 413. Szulejko, J. E.; McMahon, T. B. *J. Am. Chem. Soc.* **1993**, *115*, 7839. Smith, B. J.; Radom, L. *J. Am. Chem. Soc.* **1993**, *115*, 4885.

(4) Lias, S. G.; Bartmess, J. E.; Liebman, J. F.; Holmes, J. L.; Levin, R. D.; Mallard, W. G. *J. Phys. Chem. Ref. Data* **1988**, *17* (Suppl. 1).

(5) Bartmess, K. E.; Scott, J. A.; McIver, R. T., Jr. *J. Am. Chem. Soc.* **1979**, *101*, 6046. Graul, S. T.; Squires, R. R. *J. Am. Chem.* Soc. **1990**, *112*, 2517.

(6) Bickelhaupt, F. M. *Mass Spectrom. Rev.* **2001**, *20*, 347. Bickelhaupt, F. M.; Baerends, E. J.; Nibbering, N. M. M.; Ziegler, T. *J. Am. Chem. Soc.* **1993**, *115*, 9160. Bickelhaupt, F. M.; Buisman, G. J. H.; de Koning, L. J.; Nibbering, N. M. M.; Baerends, E. J. *J. Am. Chem. Soc.* **1995**, *117*, 9889. Bickelhaupt, F. M.; Buisman, G. J. H.; de Koning, L. J.; Nibbering, N. M. M.; Baerends, E. J. *J. Am. Chem. Soc.* **1996**, *118*, 1579. Bickelhaupt, F. M.; de Koning, L. J.; Nibbering, N. M. M. *J. Org. Chem.* **1993**, *58*, 2436.

(7) Bickelhaupt, F. M.; Baerends, E. J.; Nibbering, N. M. M. *Chem.—Eur. J.* **1996**, *2*, 196.

(8) Born, M. *Z. Phys.* **1920**, *1*, 45. Onsager, L. *J. Am. Chem. Soc.* **1936**, *58*, 1486. Wong, M. W.; Frisch, M. J.; Wiberg, K. B. *J. Am. Chem. Soc.* **1991**, *113*, 4776.

(9) Dreizler, R.; Gross, E. *Density Functional Theory*; Plenum Press: New York, 1995. Koch, W.; Holthausen, M. C. *A Chemist's Guide to Density Functional Theory*; Wiley-VCH: Weinheim, Germany, 2000. Parr, R. G.; Yang, W. *Density Functional Theory of Atoms and Molecules*; Oxford University Press: New York, 1989.

(10) Ernzerhof, M.; Perdew, J. P.; Burke, K. Density functionals: Where do they come from, why do they work? In *Topics in Current Chemistry*; Nalejawski, R. F., Ed.; Springer: Berlin, Germany, 1996; Vol. 180, pp 1. Kurth, S.; Perdew, J. P.; Blaha, P. *Int. J. Quantum Chem.* **1999**, *75*, 889. Perdew, J. P.; Ruzsinszky, A.; Tao, J. M.; Staroverov, V. N.; Scuseria, G. E.; Csonka, G. I. *J. Chem. Phys.* **2005**, *123*, 062201. Perdew, J. P.; Tao, J. M.; Staroverov, V. N.; Scuseria, G. E. *J. Chem. Phys.* **2004**, *120*, 6898.

(11) Blondel, C.; Cacciani, P.; Delsart, C.; Trainham, R. *Phys. Rev. A* **1989**, *40*, 3698. Bradforth, S. E.; Kim, E. H.; Arnold, D. W.; Neumark, D. M. *J. Chem. Phys.* **1993**, *98*, 800. Cook, P. A.; Langford, S. R.; Ashfold, M. N. R.; Dixon, R. N. *J. Chem. Phys.* **2000**, *113*, 994. Cox, J. D.; Wagman, D. D.; Medvedev, V. A. *CODATA Key Values for Thermodynamics*; Hemisphere: New York, 1989. Ellison, G. B.; Engelking, P. C.; Lineberger, W. C. *J. Am. Chem. Soc.* **1978**, *100*, 2556. Ervin, K. M.; Lineberger, W. C. *J. Phys. Chem.* **1991**, *95*, 1167. Gurvich, L. V.; Veyts, I. V.; Alcock, C. B. *Thermodynamic Properties of Individual Substances*, 4th ed.; Hemisphere Publishing Corporation: New York, 1989; Vol. 1, Parts 1−2. Hanstorp, D.; Gustafsson, M. *J. Phys. B* **1992**, *25*, 1773. Harich, S. A.; Hwang, D. W. H.; Yang, X. F.; Lin, J. J.; Yang, X. M.; Dixon, R. N. *J. Chem. Phys.* **2000**, *113*, 10073. Hepburn, J. W.; Martin, J. D. D. *Faraday Discuss.* **2000**, 416. Martin, J. D. D.; Hepburn, J. W. *J. Chem. Phys.* **1998**, *109*, 8139. Mordaunt, D. H.; Ashfold, M. N. R. *J. Chem. Phys.* **1994**, *101*, 2630. Mordaunt, D. H.; Ashfold, M. N. R.; Dixon, R. N. *J. Chem. Phys.* **1996**, *104*, 6460. Murray, K. K.; Miller, T. M.; Leopold, D. G.; Lineberger, W. C. *J. Chem. Phys.* **1986**, *84*, 2520. Ruscic, B.; Litorja, M.; Asher, R. L. *J. Phys. Chem. A* **1999**, *103*, 8625. Ruscic, B.; Litorja, M.; Asher, R. L. *J. Phys. Chem. A* **2000**, *104*, 8600. Shiell, R. C.; Hu, X. K.; Hu, Q. C. J.; Hepburn, J. W. *Faraday Discuss.* **2000**, 331. Shiell, R. C.; Hu, X. K.; Hu, Q. J.; Hepburn, J. W. *J. Phys. Chem. A* **2000**, *104*, 4339. Smith, J. R.; Kim, J. B.; Lineberger, W. C. *Phys. Rev. A* **1997**, *55*, 2036. Terentis, A. C.; Kable, S. H. *Chem. Phys. Lett.* **1996**, *258*, 626. Wickham-Jones, C. T.; Ervin, K. M.; Ellison, G. B.; Lineberger, W. C. *J. Chem. Phys.* **1989**, *91*, 2762.

(12) Ervin, K. M.; DeTuri, V. F. *J. Phys. Chem. A* **2002**, *106*, 9947.

(13) Perdew, J. P. *Phys. Rev. B* **1986**, *33*, 8822. Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098.

(14) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648. Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. *J. Phys. Chem.* **1994**, *45*, 11623.

(15) Baerends, E. J.; Autschbach, J.; Bérces, A.; Bo, C.; Boerrigter, P. M.; Cavallo, L.; Chong, D. P.; Deng, L.; Dickson, R. M.; Ellis, D. E.; Fan, L.; Fischer, T. H.; Fonseca Guerra, C.; van Gisbergen, S. J. A.; Groeneveld, J. A.; Gritsenko, O. V.; Grüning, M.; Harris, F. E.; van den Hoek, P.;

Jacobsen, H.; van Kessel, G.; Kootstra, F.; van Lenthe, E.; McCormack, D. A.; Osinga, V. P.; Patchkovskii, S.; Philipsen, P. H. T.; Post, D.; Pye, C. C.; Ravenek, W.; Ros, P.; Schipper, P. R. T.; Schreckenbach, G.; Snijders, J. G.; Solà, M.; Swart, M.; Swerhone, D.; te Velde, G.; Vernooijs, P.; Versluis, L.; Visser, O.; van Wezenbeek, E.; Wiesenekker, G.; Wolff, S. K.; Woo, T. K.; Ziegler, T. *ADF 2004.01*; SCM: Amsterdam, 2004.

(16) te Velde, G.; Bickelhaupt, F. M.; Baerends, E. J.; Fonseca Guerra, C.; van Gisbergen, S. J. A.; Snijders, J. G.; Ziegler, T. *J. Comput. Chem.* **2001**, *22*, 931.

(17) van Lenthe, E.; Baerends, E. J. *J. Comput. Chem.* **2003**, *24*, 1142.

(18) van Lenthe, E.; Baerends, E. J.; Snijders, J. G. *J. Chem. Phys.* **1993**, *99*, 4597.

(19) Vosko, S. H.; Wilk, L.; Nusair, M. *Can. J. Phys.* **1980**, *58*, 1200.

(20) Swart, M.; Snijders, J. G. *Theor. Chem. Acc.* **2003**, *110*, 34.

(21) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785. Handy, N. C.; Cohen, A. J. *Mol. Phys.* **2001**, *99*, 403.

(22) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865. Hamprecht, F. A.; Cohen, A. J.; Tozer, D. J.; Handy, N. C. *J. Chem. Phys.* **1998**, *109*, 6264. Boese, A. D.; Handy, N. C. *J. Chem. Phys.* **2001**, *114*, 5497.

(23) Van Voorhis, T.; Scuseria, G. *J. Chem. Phys.* **1998**, *109*, 400. Proynov, E. I.; Sirois, S.; Salahub, D. R. *Int. J. Quantum Chem.* **1997**, *64*, 427.

(24) Tao, J. M.; Perdew, J. P.; Staroverov, V. N.; Scuseria, G. E. *Phys. Rev. Lett.* **2003**, *91*, 146401. Boese, A. D.; Handy, N. C. *J. Chem. Phys.* **2002**, *116*, 9559.

(25) Perdew, J. P.; Ernzerhof, M.; Burke, K. *J. Chem. Phys.* **1996**, *105*, 9982. Becke, A. D. *J. Chem. Phys.* **1997**, *107*, 8554. Adamo, C.; Barone, V. *Chem. Phys. Lett.* **1997**, *274*, 242.

(26) Jensen, F. *Introduction to computational chemistry*; Wiley & Sons: Chichester, U. K., 1999.

(27) Bickelhaupt, F. M.; Hermann, H. L.; Boche, G. *Angew. Chem., Int. Ed.* **2005**, *44*, in press.

CT0502460

JCTC Journal of Chemical Theory and Computation

Quantum Mechanical Calculations for Benzene Dimer Energies: Present Problems and Future Challenges

W. Bernd Schweizer* and Jack D. Dunitz*

Organic Chemistry Laboratory, Swiss Federal Institute of Technology, ETH-Hönggerberg, CH-8093 Zurich, Switzerland

Received September 20, 2005

Abstract: Factors influencing quantum mechanical calculations of nonbonded interactions between organic molecules are still imperfectly understood. Much effort has gone into efforts to calculate the structures and binding energies of stable benzene dimers. However, little experimental evidence is available for comparison with theoretical results. As a benchmark for assessing the reliability and accuracy of such calculations, the benzene crystal structure seems a more suitable target than the elusive dimer structures.

For some time now, quantum mechanical calculations have been providing reliable answers to many questions about the binding energies, atomic arrangements, electron density distributions, electrical moments, and vibrational frequencies of small- to medium-sized molecules. Computer programs for carrying out the necessary computations at several theoretical levels are readily available, and many of the calculations can now be carried out with a desktop computer. During the past 10 years or so, attempts have been made to extend such calculations to questions of intermolecular binding. Since intermolecular interaction energies are only a small fraction of intramolecular bond energies, reliable answers to such questions are much more difficult to obtain.

As an example, take benzene. Although several quantum mechanical studies of the preferred structures and binding energies of benzene dimers[1−9] have been made, there are still serious outstanding problems. Concerning the preferred structures, calculations agree that the T-shaped and parallel-displaced (PD) dimers are the most stable, with approximately equal energies, and that the energy hypersurface is rather flat with a low interconversion barrier. There is less agreement about the binding energies of the dimers. Since the intermolecular attractions are largely due to London dispersion effects, they cannot be adequately handled by Kohn−Sham density functional theory.[10] Additionally, dispersion energies are not taken into account at the Hartree−Fock level, which treats the interaction of each electron with the averaged distribution of the other electrons. Thus, at the Hartree−Fock level, enlarging the basis set does not have much influence on the binding energies of dimers, but it has a very big effect on MP2 energies, which tend to overestimate the stabilization of the dimer because of the so-called basis-set superposition error (BSSE), as judged from results of higher-level calculations.

According to one recent study,[6] high-level calculations with different basis sets and different methods of allowance for electron correlation and basis-set superposition error yield binding energies for the T-shaped dimer ranging from 1.40 to 3.63 kcal mol^{-1}, with 2.7 kcal mol^{-1} as the preferred value. Corresponding energies for the PD dimer range from 2.02 to 4.95 kcal mol^{-1}, with 2.8 kcal mol^{-1} as the preferred value. Even though the preferred values are probably close to the correct binding energies (by an elaborate procedure, Tsuzuki et al.[4] obtain similar binding energies−2.46 and 2.48 kcal mol^{-1}, respectively, for the two benzene dimers), this is a discouraging result. An uncertainty on the order of a kcal mol^{-1} is too large to provide reliable answers to the problems of interest. Experimental evidence about the structures and binding energies of benzene dimers and small clusters is scarce and difficult to interpret.[2] Diverse studies using different techniques for preparing and analyzing gas-phase clusters have not yielded consistent structures or energies for benzene dimers. Neutron diffraction shows that there are

* Corresponding author phone: +41 44 632 45 07; e-mail: schweizer@org.chem.ethz.ch.

10.1021/ct0502357 CCC: $33.50
Published on Web 12/14/2005

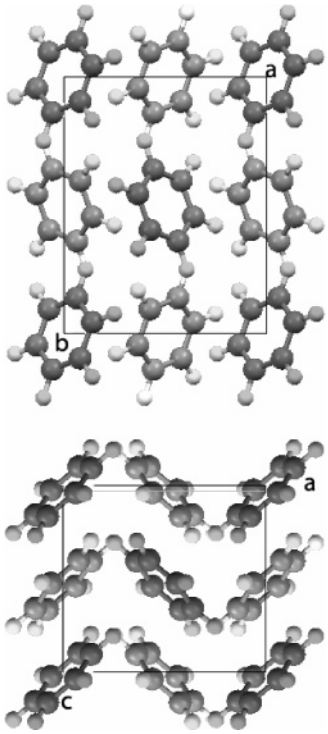


Figure 1. Crystal structure of benzene, viewed down the *b* and *c* axes of the unit cell.

no preferred orientations of neighboring molecules in liquid benzene.[11] Thus, there are really no reliable experimental data by which to assess and compare the results of the various computational studies. In place of the benzene dimers, a more suitable target would seem to be the calculation of the lattice energy of crystalline benzene. This is an experimental quantity whose value is known within reasonable limits. A calculation of the energy of the benzene crystal by quantum mechanical methods might first appear to be a much more formidable task than that of calculating the binding energies of benzene dimers, but with a few obvious simplifications and approximations, it should be perfectly feasible. It is, at least, a challenge.

The crystal structure of benzene has been the subject of countless experimental and theoretical studies. At normal pressures, benzene crystallizes in the space group *Pbca* with unit cell dimensions $a = 7.39$, $b = 9.42$, and $c = 6.81$ Å at 138 K.[12] The atomic arrangement is shown in Figure 1. The heat of sublimation of benzene has been variously measured as between 40 and 45 kJ mol^{-1},[13] with a preferred value of 44.4 kJ mol^{-1}. The switch from kcal mol^{-1} to kJ mol^{-1} energy units should be noted here. The lattice energy of a crystal can be derived as the sum of interaction energies between a central reference molecule and all the other molecules (divided by 2 as result of the counting method).[14] Since the interaction energy falls off rapidly with increasing intermolecular separation, the sum is essentially limited to the contributions of the 12 or 14 first neighbors of the reference molecule, the first coordination shell. Expansion of the shell to include contributions from more distant partners usually adds only a few percent to the sum. For molecular crystals, the error introduced by ignoring many-body effects in such a straightforward summation is not serious. For the benzene crystal structure, with its high symmetry, we need to consider only four types of molecular neighbor pairs, illustrated in Figure 2, with approximate interaction energies in Table 1. These pairs are *not* minimum energy pairs. They are compromises, subject to the pulls and pushes of neighboring molecules in the repeating crystal pattern, but they are, so to say, the building blocks of the crystal. Their energies were estimated by the semiclassical density sums or Pixel method[15−18] and at two ab initio computational levels using 6-31+g(d) and 6-31++g(d,p) basis sets of orbitals, including second-order Møller−Plesset correlation energy (MP2 treatment), both with and without counterpoise corrections (Gaussian 03).[19] The calculations were based on the experimental atomic coordinates and unit cell dimensions[12,20] without any energy minimization, so that the benzene molecules in the computational scheme deviate slightly from their ideal D_{6h} symmetry. These small deviations from ideality should not have a serious effect on the pattern of pair energies.

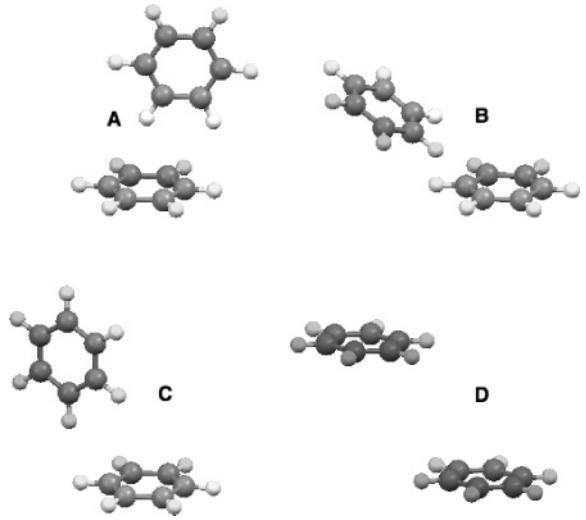


Figure 2. Molecular pairs A, B, C, and D involved in the first coordination shell of a given molecule in the crystal structure of benzene. The pairs A, B, and C are produced by the glide-reflection symmetry operations of the space group, the pair D by the *c* translation (see Table 1).

As seen from Figure 2, a given reference molecule is engaged in four pairs of each type A, B, and C (glide reflections) and in two pairs of type D ($\pm c$ axis translation). Taking the sum of the calculated interaction energies with the appropriate multiplicities and dividing by 2, we obtain a value of 43.8 kJ mol^{-1} for the estimated lattice energy with the Pixel values, as good an agreement with the experimental value as we could possibly hope for. The results of the corresponding ab initio calculations are not so clear. In concordance with earlier experience,[3,6] the MP2 calculations greatly overestimate intermolecular binding and give estimated lattice energies of 89 and 111 kJ mol^{-1}, which are at least double the experimental value. Counterpoise corrections indeed reduce these energies to less than half of the uncorrected values, but the results still miss the mark. Compared with the experimental value, the smaller basis set yields too low a lattice energy (34 kJ mol^{-1}), the larger basis

Table 1. Interaction Energies (kJ mol^{-1}) of Molecular Pairs A−D (Figure 2) in the *Pbca* Crystal Structure of Benzene, as Calculated by the Pixel Method [E_{PIX}, Electron Density Calculated with MP2/6-31++g(d,p)] and by ab Initio Calculations with Two Orbital Basis Sets [6-31+g(d) and 6-31++g(d,p)] at the MP2 Level without ($E_{MP2}{}^{+}$ and $E_{MP2}{}^{++}$) and with Counterpoise Correction ($E_{CP}{}^{+}$ and $E_{CP}{}^{++}$)[a]

pair	symmetry operation	*N*	E_{PIX}	$E_{MP2}{}^{+}$	$E_{CP}{}^{+}$	$E_{MP2}{}^{++}$	$E_{CP}{}^{++}$	*D* (Å)
A	*a*/*c* glide reflection	4	−9.0	−19.6	−7.9	−23.0	−10.5	5.02
B	*c*/*b* glide reflection	4	−6.1	−12.1	−4.8	−15.5	− 7.3	5.81
C	*b*/*a* glide reflection	4	−5.1	−10.6	−3.5	−13.7	− 6.1	5.99
D	±*c* translation	2	−1.7	− 4.5	−2.0	− 6.9	− 4.0	6.81
	lattice energy estimate		−43.8	−89.1	−34.4	−111.3	−51.8	

[a] The distances *D* are between centers of mass of the two molecules in the pair. Column *N* gives the number of symmetry-related pairs involving a given reference molecule.

Table 2. Pixel Energies (kJ mol^{-1}) for the Molecular Pairs A−D (Figure 2) Based on Charge Densities Calculated at Different Theoretical Levels[a]

pair	symmetry operation	*N*	HF/3-21G	MP2/6-31G**	MP2/6-31+G(d)	MP2/6-31++G(d,p)
A	*a*/*c* glide reflection	4	−14.9	−11.2	−9.3	−9.0
B	*c*/*b* glide reflection	4	−10.2	−7.7	−7.7	−6.1
C	*b*/*a* glide reflection	4	−7.8	−6.2	−5.1	−5.1
D	±*c* translation	2	−2.0	−1.6	−1.8	−1.7
	lattice energy estimate		−69.8	−53.4	−47.8	−43.8

[a] Column *N* gives the number of symmetry-related pairs involving a given reference molecule.

set too large an energy (52 kJ mol^{-1}). The simple Pixel calculation gives the best result, and it is interesting that the individual pair energies quite closely parallel those from the MP2/6-31++g(d,p) calculation with counterpoise correction ($E_{CP}{}^{++}$).

The experimental estimate of the lattice energy of benzene (40−45 kJ mol^{-1}) can be regarded as a kind of benchmark for assessing the reliability of high-quality quantum mechanical calculations for benzene dimers. The Pixel calculation comes close to the mark, but its partitioning of the energy into Coulombic, polarization, dispersion, and repulsion contributions is to some extent arbitrary and parameter-dependent.[15−17] The MP2 calculations give more than double the correct value and are clearly not very useful. The counterpoise-corrected values yield erratic values for the lattice energy, and it is not obvious how any larger basis set or improved BSSE correction would influence the result. In contrast to the behavior of the quantum mechanical pair energies, Pixel energies show a steady decrease as the quality of the theoretical level of the calculation improves (Table 2).[20] The low-level Hartree−Fock 3-21g charge density is wide of the mark, while the final value (MP2/6-31++g(d,p)) is within the experimental range.

In principle, interaction energies of the molecular pairs A−D are no more difficult to calculate than those of the T-shaped and PD dimers —easier, in fact, since no energy minimization procedures are needed—but with a present uncertainty on the order of 1 kcal mol^{-1} (∼4 kJ mol^{-1}) in the estimated interaction energy of each benzene molecular pair, as might be inferred from the range of results for the benzene dimers, the goal of calculating an unconditionally reliable ab initio value for the lattice energy of benzene may still seem remote, but it is on the horizon and should be attainable.

References

(1) Jaffe, R. L.; Smith, G. D. *J. Chem. Phys.* **1996**, *105*, 2780−2788.

(2) Smith, G. D.; Jaffe, R. L. *J. Phys. Chem.* **1996**, *100*, 9624−9630.

(3) Hobza, P.; Selzle, H. L.; Schlag, E. W. *J. Phys. Chem.* **1996**, *100*, 18790−18794.

(4) Tsuzuki, S.; Honda, K.; Uchimaru, T.; Mikami, M.; Tanabe, K. *J. Am. Chem. Soc.* **2002**, *124*, 104−112.

(5) Tsuzuki, S.; Uchimaru, T.; Sugawara, K.; Mikami, M. *J. Chem. Phys.* **2002**, *117*, 11216−11220.

(6) Sinnokrot, M. O.; Valeev, E. F.; Sherrill, C. D. *J. Am. Chem. Soc.* **2002**, *124*, 10887−10893.

(7) Reyes, A.; Tlenkopatchev, M. A.; Fomina, L.; Guadarrama, P.; Fomine, S. *J. Phys. Chem. A* **2003**, *107*, 7027−7031.

(8) Ye, X.; Li, Z.-H.; Wang, W.; Fan, K.; Xu, W.; Hua, Z. *Chem. Phys. Lett.* **2004**, *397*, 56−61.

(9) Zhikol, O. A.; Shishkin, O. V.; Lyssenko, K. A.; Leszczynski, J. *J. Chem. Phys.* **2005**, *122*, 144104 1−8.

(10) Van Mourik, T.; Gdanitz, R. J. *J. Chem. Phys.* **2002**, *116*, 9620−9623.

(11) Cabaço, M. I.; Danten, Y.; Besnard, M.; Guissani, Y.; Guillot, B. S. *J. Phys. Chem. B* **1997**, *101*, 6977−6987.

(12) Bacon, G. E.; Curry, N. A.; Wilson, S. A. *Proc. R. Soc. London, Ser. A* **1964**, *279*, 98−110.

(13) For a recent compilation of experimental sublimation enthalpies, see Chickos, W. E.; Acree, W. E., Jr. *J. Phys. Chem. Ref. Data* **2002**, *31*, 537−698.

(14) Why divide by two? This question may occur to readers unfamiliar with the bookkeeping of lattice energy. For simplicity, consider a cluster of *N* monatomic molecules. Each molecule interacts with all the others, so there are $N(N-1)/2$ separate interactions, each of which contributes to the total cohesive energy of the cluster. If this energy is to be expressed in molar units, e.g., kJ mol^{-1}, then the sum

over the cluster has to be divided by N, the number of molecules in the cluster. For an extended crystal, where all molecules can be regarded as equivalent, the same result is obtained by choosing an arbitrary reference molecule, summing over the $N - 1$ interactions with the remaining molecules, and then dividing by 2. If the molecules in the crystal are not equivalent, e.g., if $Z' > 1$ or in a cocrystal, then the matter becomes more complicated.

(15) Gavezzotti, A. *J. Phys. Chem. B* **2002**, *106*, 4145−4154.

(16) Gavezzotti, A. *J. Phys. Chem. B* **2003**, *107*, 2344−2353.

(17) Gavezzotti, A. *J. Chem. Theory Comput*. **2005**, *1*, 834−840.

(18) Dunitz, J. D.; Gavezzotti, A. *Angew. Chem., Int. Ed*. **2005**, *44*, 1766−1787.

(19) 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.; W. Gill, P. M.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03*, revision C.02; Gaussian, Inc.: Wallingford, CT, 2004.

(20) This information can be conveniently recovered from the Cambridge Structural Database (CSD), distributed by the Cambridge Crystallographic Data Centre, Cambridge, England (www.ccdc.cam.ac.uk), under the refcode BENZEN01.

(21) In the four Pixel calculations (Table 2), the dispersion contribution to the binding energy is dominant and remains nearly constant. The largest energy changes occur in the repulsion contribution, which increases with the expansion of the basis set.

CT0502357

JCTC Journal of Chemical Theory and Computation

Electronic Excitations of the Chromophore from the Fluorescent Protein asFP595 in Solutions

Alexander V. Nemukhin,*[,†,‡] Igor A. Topol,[§] and Stanley K. Burt[§]

Department of Chemistry, M. V. Lomonosov Moscow State University, Moscow, 119992, Russia, Institute of Biochemical Physics, Russian Academy of Sciences, Moscow, 119997, Russia, and Advanced Biomedical Computing Center, National Cancer Institute at Frederick, Frederick, Maryland 21702

Received September 28, 2005

Abstract: We present the results of modeling spectral properties of the chromophore, 2-acetyl-4-(*p*-hydroxybenzylidene)-1-methyl-5-imidazolone (AHBMI), from the newly discovered fluorescent protein asFP595 in different solvents and compare computational and recent experimental data. The time-dependent density functional theory (TDDFT) method is used to estimate positions of spectral bands with large oscillator strengths for vertical transitions to excited states following geometry optimizations of chromophore coordinates in vacuo and in solutions. The performance of different TDDFT functionals in computing excitations for a simpler chromophore from the green fluorescent protein was tested at the preliminary stage. Properties of various protonation states (neutral, anionic, zwitterionic) for the cis and trans conformations of AHBMI are compared. By using the polarizable continuum model, the following solvents have been considered for AHBMI: water, ethanol, acetonitrile, and dimethyl sulfoxide. It is shown that the bands found experimentally in aqueous solution refer to the cis neutral and cis anionic (or trans zwitterionic) conformations. The computed band positions deviate from experimental ones in water by no more than 35 nm (0.23 eV). In accord with experimental studies, the band shifts in different solvents do not show correlation with the dielectric constant or dipole moment; however, the computed values of the shifts are much smaller than those measured experimentally for the ionic species.

Introduction

Proteins from the family of the green fluorescent proteins (GFP) are extensively used in molecular and cell biology[1−3] and promise a variety of important biotechnology applications.[4] A newly discovered GFP-like protein from the sea anemone *Anemonia sulcata* asFP595[5] is initially nonfluorescent, but in response to intense green light irradiation at 568 nm, it becomes brightly fluorescent (kindles) with emission at 595 nm. Photoswitching properties of this kindling fluorescent protein may be useful for information storage in macromolecules or for creating triggerable markers in living cells. The mechanism of kindling is far from clear, and at present, substantial efforts are being undertaken to understand the intriguing properties of asFP595.[6−9]

The model chromophore from the kindling protein, 2-acetyl-4-(*p*-hydroxybenzylidene)-1-methyl-5-imidazolone (AHBMI), was recently elegantly synthesized[9] following crystallographic studies of the chromoprotein asFP595.[6] Its spectral properties in solution and their dependence on the pH and polarity of the solvent were investigated.[9] It was suggested that the bands in aqueous solution at 418 nm (2.97 eV) and 520 nm (2.38 eV) referred to the neutral and anionic states of the model chromophore. These absorption maxima experienced noticeable shifts in ethanol, 2-propanol, and

* Corresponding author phone: 7-095-939-1096; fax: 7-095-939-0283; e-mail: anem@lcc.chem.msu.ru.
† M. V. Lomonosov Moscow State University.
‡ Russian Academy of Sciences.
§ National Cancer Institute at Frederick.

10.1021/ct050243n CCC: $33.50
Published on Web 12/23/2005

dimethylformamide compared to water, although no clear correlation with the dielectric constant of the solvent was seen.

The goal of this work is to model spectral features of the vertical S_0-S_1 transitions of AHBMI in vacuo and in solutions by using the time-dependent density functional theory (TDDFT) method[10,11] in conjunction with the polarizable continuum model (PCM)[12,13] for solvents.

Since the pioneering estimates of the S_0-S_1 excitation energies in the GFP chromophore at the semiempirical INDO/S level by Voityuk and coauthors,[14] numerous attempts to calculate optical spectra of the GFP-like chromophores have been described in the literature. The reviews of Helms[15] and recent publications of Das et al.,[16] Marques et al.,[17] Laino et al.,[18] Toniolo et al.,[19] Martin et al.,[20] Vendrell et al.,[21] Altoe et al.,[22] Sinicropi et al.,[23] and Lopez et al.[24] present achievements of the theory in this field. Presently, it is believed that either the CASPT2//CASSCF[20−23] or TDDFT[17,21,24] method may provide reasonable excitation energies, although accurate prediction of the spectral band positions is still a problem. The state-of-the-art CASPT2//CASSCF approach was found to reproduce the absorption wavelength with a less than 40 nm error.[23]

The vast majority of theoretical simulations have been carried out for the simplest model GFP chromophores, for example, 4-hydroxybenzylidene-1,2-dimethylimidazolinone (HBDI).[15−24] The larger chromophore from asFP595, AHBMI, closely resembles the chromophores from the red fluorescent proteins DsRed and HcRed, whose spectral properties have been modeled in vacuo by using the TDDFT method in the B3LYP/6-31++G(d,p)//B3LYP/6-31+G(d,p) approximation.[25,26] The TDDFT technique, also at the same level, was applied by Xie and Zeng[27] for the characterization of another closely related chromophore, 4′-hydroxybenzylidene-2-methyl-imidazolin-5-one-3-acetate (HBMIA), in various protonation states of cis and trans isomers in vacuo followed by estimates of solvent effects in aqueous solution. For all these simulations, there are some quantitative discrepancies between computed and experimental absorption bands either in chromoproteins[25,26] or in solutions.[27]

In the following sections, we describe, first, the calculations of excitation energies for the anionic form of the simplest GFP model chromophore by exploring different TDDFT functionals and, second, the TDDFT calculations for the asFP595 chromophore in the B3LYP/6-311++G(2df,p)//B3LYP/6-31+G(d,p) approximation. In the latter case, optimization of the geometry parameters of all species and estimates of the spectral properties have been performed for gas-phase and solvent environments (water, ethanol, acetonitrile, and dimethyl sulfoxide) within the polarizable solvation model. These data are compared to the measurements of the absorption bands for this chromophore in water, ethanol, and dimethylformamide.[9]

Results for the GFP Chromophore

Different estimates of the possible errors of the TDDFT approximation for the GFP-like chromophores may be found in the literature. For vertical excitation energies of the S_0-S_1 transitions, these range from 0.44[21] to 0.1 eV.[24] Therefore, we performed preliminary calculations for the simplest model system relying on the available gas-phase experimental results for the excitation energy which refers to the GFP anionic state: 2.59 eV (479 nm).[28] Calculations have been carried out with conventional options of Gaussian 03.[29,10] Figure 1 illustrates equilibrium geometry configuration of the cis anionic form of the GFP chromophore. The coordinates of the molecule have been optimized by using the B3LYP/6-31++G(d,p) method.

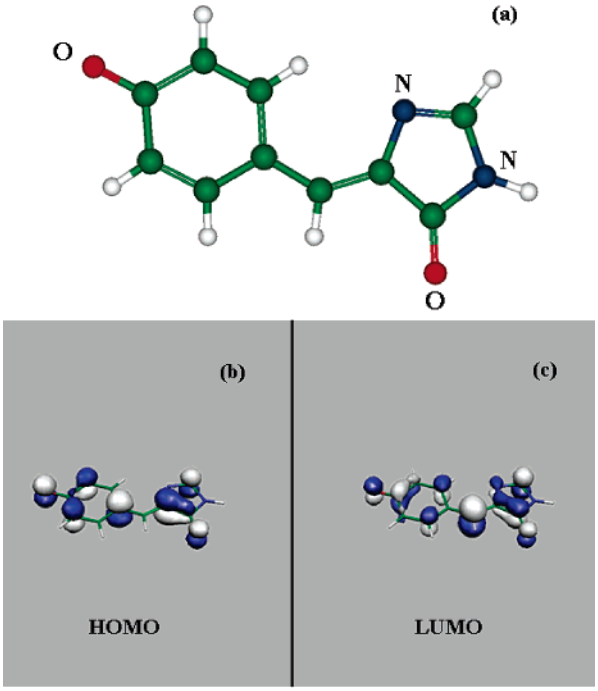


Figure 1. Structure of the GFP chromophore (a) and views of the HOMO (b) and LUMO (c) orbitals computed in the B3LYP approximation.

In Table 1, we collect computed properties of the vertical excitations of the GFP anion calculated in different approximations. Excitation energies, corresponding wavelengths, and oscillator strengths are presented. The first part of Table 1 shows the dependence of computed parameters on the type of exchange-correlation functional in TDDFT while retaining the same basis set 6-31++G(d,p). The second part of Table 1 illustrates the basis set dependence for B3LYP as a choice of the particular functional in TDDFT. The dominant contribution to this $\pi \rightarrow \pi^*$ type excitation refers to the HOMO → LUMO transition; however, some mixtures from other orbitals appear for nonhybrid functionals. The sequence number (first, second, or third) of the needed excitation in every case is easily recognized by the value of the oscillator strength; other states (even with lower excitation energies) are characterized by *f* values close to zero. All the states with low oscillator strengths below the bright state are of the same $\pi \rightarrow \pi^*$ origin.

As follows from these simulations, all local-density approximation (LDA) and gradient-corrected DFT functionals give very similar results. Hybrid functionals also provide values of the same quality; however, the more Hartree−Fock (HF) exchange is included in the functional, the greater is

Table 1. Characteristics of the Vertical Excitations of the GFP Anion Calculated in Different Approximations[a]

variables	excitation number	energy of HOMO (au)	ΔE (eV)	λ (nm)	oscillator strength, f
Different Functionals (FUN) in FUN/6-31++G(d,p)// B3LYP/6-31++G(d,p) Calculations					
RHF	1	−0.114	3.50	353.8	1.206
SVWN[b]	2	−0.060	2.94	421.1	0.788
BVWN[c]	3	−0.060	2.94	421.9	0.798
BP86[d]	3	−0.037	2.94	421.1	0.795
VSXC[e]	3	−0.033	3.02	410.1	0.834
MPW1PW91[f]	2	−0.061	3.11	399.1	0.933
B972[g]	2	−0.053	3.09	400.9	0.919
B1LYP[h]	2	−0.054	3.08	402.7	0.924
B3LYP[i]	2	−0.055	3.05	406.0	0.902
BHandHLYP[j]	1	−0.082	3.23	384.4	1.038
Different Basis Sets (BS) in B3LYP/BS// B3LYP/6-31++G(d,p) Calculations					
6-31G	2	−0.036	3.25	381.2	0.930
6-31G(d)	2	−0.032	3.22	384.5	0.899
6-31+G(d)	1	−0.054	3.06	404.8	0.906
6-31++G(d,p)	2	−0.055	3.05	406.0	0.902
6-31++G(2df,p)	2	−0.056	3.03	408.8	0.880

[a] In all cases, geometry coordinates have been computed in the B3LYP/6-31++G(d,p) approximation. The first part of the table shows the dependence of computed parameters on the type of functional in TDDFT while retaining the same basis set 6-31++G(d,p). The notation of functionals is given according to the Gaussian 03 system.[29] The second part of the table illustrates basis set dependence for B3LYP as a choice of the exchange-correlation functional in TDDFT. [b] SVWN: (S) Slater exchange (Slater, J. C. *Quantum Theory of Molecules and Solids*; McGraw−Hill: New York, 1974; Vol. 4), (VWN) Vosko, Wilk, and Nusair correlation (Vosko, S. H.; Wilk, L.; Nusair, M. *Can. J. Phys.* **1980**, *58*, 1200). [c] BVWN: (B) Becke's 1988 exchange functional (Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098) and VWN correlation functional. [d] BP86: Becke's 1988 exchange, (P86) Perdew correlation functional (Perdew, J. P. *Phys. Rev. B* **1986**, *33*, 8822). [e] VXSC: van Voorhis and Scuseria's gradient corrected functional (Van Voorhis, T.; Scuseria, G. E. *J. Chem. Phys.* **1998**, *109*, 400). [f] MPW1PW91: Modified Perdew−Wang exchange and Perdew−Wang 91 correlation functionals (Adamo, C.; Barone, V. *J. Chem. Phys.* **1998**, *108*, 664). [g] B972: Wilson, Bradley, and Tozer's modification to the initial B971 functional (Orig.: Hamprecht, F. A.; Cohen, A. S. J.; Tozer, D. J.; Handy, N. C. *J. Chem. Phys.* **1998**, *109*, 6264. Modif.: Wilson, P. J.; Bradley, T. J.; Tozer, D. J. *J. Chem. Phys.* **2001**, *115*, 9233). [h] B1LYP: (B1) Becke's one-parameter hybrid exchange functional (Becke, A. D. *J. Chem. Phys.* **1996**, *104*, 1040); (LYP) Lee, Yang, and Parr correlation functional (Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785). [i] B3LYP: (B3) Becke's three-parameter hybrid exchange functional (Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648); LYP correlation. [j] BHandHLYP: (BHandH) Half-and-half hybrid exchange functional [$0.5E_x$(HF) + $0.5E_x$(LSDA) + $0.5x\Delta E_x$(Becke88)]; LYP correlation.

the energy gap ΔE and the deviations from experiment are larger. The basis set dependence is saturated fairly fast.

A somehow discouraging result is that, in all approximations, the deviations from the experimental parameters (2.59 eV, 479 nm)[28] are considerable, giving rise to at least a 0.35 eV error. The best performance of the B3LYP method is characterized by the 0.44 eV error. These error bars seem to be typical for TDDFT applications to such complex systems.

Results for the Gas-Phase and Aqueous Structures of the Chromophore from asFP595

The panels of Figure 2 show equilibrium geometry configurations of cis and trans isomers of the neutral, anionic, and zwitterionic forms of the chromophore AHBMI. We do not overcrowd the pictures with the computed geometry parameters, which have been reported in many papers describing the GFP-like chromophores.[2] Instead, we collect the Cartesian coordinates of the corresponding structures optimized for the gas-phase conditions in the B3LYP/6-31+G(d,p) approximation in the Supporting Information. We should mention that the only nonplanar structure among those considered here refers to the cis zwitterionic species. This is due to the repulsion of the nearby hydrogen atoms from the five- and six-member rings, occurring in this particular arrangement. Reoptimization of the geometry parameters in the dielectric continuum corresponding to the aqueous solution ($\epsilon = 80$) leads to small changes by no more than 0.02 Å in bond lengths and 1° in angles.

In Table 2, we present the total energies in vacuo and free energies in aqueous solution and the relative (trans vs cis) energies for these structures calculated in the B3LYP/6-31+G(d,p) approximation. According to these data, cis conformations are preferable for all species except the zwitterionic structure in aqueous solution.

A widely cited contribution to the question of the cis−trans isomerization of the GFP-like chromophore in the ground electronic state was the study of He et al.[30] These authors used NMR spectroscopy to characterize conformations of HBDI in water for the neutral, cationic, and anionic states. They found that, for the model chromophore, the cis isomers must be lower in energy by 0.8, 2.1, and 2.3 kcal/mol for the cationic, neutral, and anionic states, respectively. The corresponding activation barriers for cis−trans isomerization were estimated as 11.7, 13.1, and 13.1 kcal/mol for the cationic, neutral, and anionic forms, respectively.[30]

Recent quantum chemical calculations also provide support to a somewhat greater stability of cis isomers. In the semiempirical quantum mechanical/molecular mechanical calculations of Toniolo et al.[31] for the GFP chromophore in vacuo and inside the shell of explicit water molecules, it was observed that the cis conformer was more stable than trans as a consequence of "the larger dipole moment in the cis conformer". Wilmann et al.[26] presented, along with other findings for the HcRed fluorescent protein, the results of quantum chemical modeling of the cis and trans conformations of the HcRed chromophore in vacuo. According to the B3LYP/6-31++G(d,p)//B3LYP/6-31+G(d,p) calculations, the coplanar cis conformer is 1.7 kcal/mol lower in energy than the coplanar trans isomer.[26] The most recent work of Xie and Zeng[27] reported lower energies of the cis conformations for all protonation states (neutral, anionic, and zwitterionic) of the HBMIA chromophore, as computed in the B3LYP/6-31++G**//B3LYP/6-31+G** approximation.

Therefore, our results are in line with previous findings showing a slightly lower energy of the cis forms for neutral and anionic species.

In Table 3, we present the results of calculations of vertical excitation energies, corresponding wavelengths, and oscillator strengths computed in the B3LYP/6-311++G(2df,p)//B3LYP/6-31+G(d,p) approximation both in vacuo and in aqueous solution. For the solution, geometry parameters have been

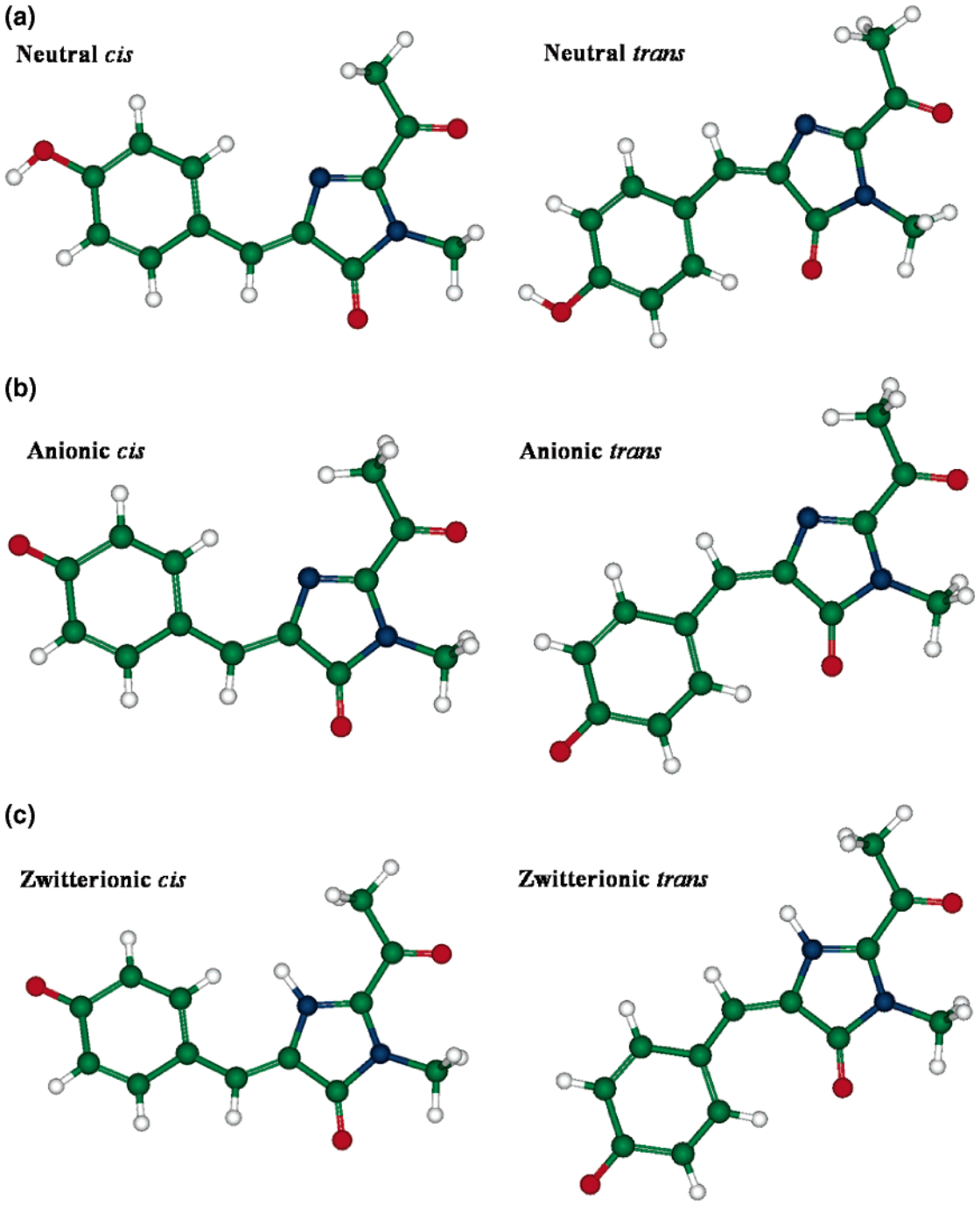


Figure 2. Structures of the AHBMI chromophore: (a) neutral, (b) anionic, (c) zwitterionic forms.

Table 2. Energies in Vacuo and Free Energies in Aqueous Solution (au) of the Forms of AHBMI (Figure 2)[a]

species	neutral	anionic	zwitterionic
	Gas Phase		
cis	−837.843 79	−837.327 00	−837.812 65
trans	−837.841 14	−837.324 06	−837.811 78
energy of trans vs cis	1.66	1.84	0.54
	Aqueous Solution		
cis	−837.849 76	−837.383 38	−837.826 37
trans	−837.847 15	−837.381 36	−837.832 82
energy of trans vs cis	1.64	1.27	−4.04

[a] Relative energies are given in kcal/mol.

optimized by using the PCM method. These data show noticeable solvent-induced shifts in band positions: −0.15 eV (+23 nm) for cis neutral, −0.09 eV (+18 nm) for cis anion, and −0.07 eV (+14 nm) for cis zwitterion. In accord with experimental findings,[9] the intensities (oscillator strengths) of bands assigned to the neutral chromophore are lower than those of the ionic species.

Comparison with Experimental Results for Excitations in Solution for the AHBMI Chromophore

The calculation results shown in Tables 2 and 3 for the aqueous solution allow us to confirm the experimental assignment that the band at a lower wavelength (418 nm) refers to the neutral chromophore and to suggest the cis neutral form (453 nm) as a primary candidate. Although the computed value for the trans isomer (437 nm) is closer to the experimental measurements, it is difficult to ignore the

Table 3. Excitation Energies, Corresponding Wavelengths, and Oscillator Strengths Computed in the B3LYP/6-311++G(2df,p)//B3LYP/6-31+G(d,p) Approximation

structure	ΔE, eV	λ, nm	oscillator strength f
Gas Phase (Geometry Optimized in Vacuo)			
cis neutral	2.89	430	0.53
cis anionic	2.56	484	0.80
cis zwitterionic	2.38	521	0.68
Aqueous Solution (Geometry Optimized in Solution)			
cis neutral	2.74	453	0.72
cis anionic	2.47	502	0.94
cis zwitterionic	2.31	535	0.85
trans neutral	2.84	437	0.82
trans anionic	2.61	474	1.18
trans zwitterionic	2.48	500	1.10

higher energy of the trans structure (1.6 kcal/mol) and a possible fairly large rotational barrier for cis−trans isomerization in solution.[30] Even for the cis form, the discrepancy between calculated and experimental values (35 nm or 0.23 eV) falls within accepted error bars of the TDDFT model, as discussed above.

A comparison of computational and experimental results for anion species appears even more encouraging. The experimental band in water (520 nm) may be assigned to either the cis anionic (502 nm) or trans zwitterionic (500 nm) forms of the chromophore. In both cases, the deviations from experimental band positions are fairly small (18 nm or 0.09 eV and 20 nm or 0.10 eV, respectively). The cis anionic structure seems more preferable, considering that this band appears experimentally at basic pH.[9]

For different solvents, we reoptimized geometry configurations of the chromophore by using the B3LYP(6-31+G(d,p) approximation and the PCM model. There is no option in Gaussian 03 to treat dimethylformamide (DMF). Therefore, we considered acetonitrile and dimethylsulfoxide (DMSO), whose dielectric constants ϵ and dipole moments μ bracketed those of DMF. Then, excitation energies, wavelengths, and oscillator strengths were computed in the B3LYP/6-311++G(2df,p) approximation. The results are presented in Table 4.

For the band associated with the neutral form in water (418 nm), an agreement between our theoretical estimates and experimental results[9] is reasonable: in both cases, the band position is predicted to be slightly sensitive to the solvent. For the band assigned to the anionic form, the distinctions are much larger, and the observed shifts of up to 50 nm when moving from water to DMF are not reproduced computationally.

We notice that the computed wavelength for the AHBMI chromophore in DMSO in the cis neutral form (458 nm) agrees perfectly with the measurements for the related chromophore HBMIA in DMSO (460 nm).[32]

Discussion and Conclusion

The application of the TDDFT method for estimates of excited state parameters for fairly large molecules including the fluorescent protein chromophores is becoming very popular. The results of simulations described in this paper contribute to this growing field of computational chemistry. To some extent, the data collected in Table 1 confirm the observation formulated in ref 24 that the use of LDA-based functionals could lead to somewhat better agreement with experimental results; however, we cannot achieve such small errors in excitation energies of 0.1 eV as reported in ref 24 for the chromophore of the blue fluorescent protein. The 40 nm deviations in band positions for the GFP chromophore (or 5 kcal/mol as reported in ref 23, which equals 0.22 eV), illustrate the efficiency of the CASPT2//CASSCF approach. The wavelengths at 402 nm reported recently by Xie and Zeng[27] for the HBMIA chromophore (following their results of TDDFT calculations and estimates of aqueous shifts) show noticeable deviations from the experimental band position of this chromophore in DMSO at 460 nm. Most likely, the errors 0.2−0.4 eV may be expected when computing excited-state energies of GFP-like chromophores.

Within these error bars, our TDDFT simulations of spectral bands of the asFP595 chromophore in water are consistent with experimental findings.[9] In simulations, we can expand the experimental knowledge and assign particular conformations to the molecule in different protonation states. From energy considerations, the neutral form should correspond to the cis isomer (consistent with the results of NMR studies[30] of the GFP chromophore), and its absorption band is computed to be at 453 nm (2.74 eV). The deviation from

Table 4. Comparison of Calculated (B3LYP/6-311++G(2df,p)//B3LYP(6-31+G(d,p)) Excitation Energies and the Corresponding Wavelengths for All Considered Solvents[a]

	neutral		anionic		zwitterionic	
solvent	ΔE, eV	λ, nm	ΔE, eV	λ, nm	ΔE, eV	λ, nm
Cis Isomers						
gas phase ($\epsilon = 1$)	2.89	430	2.56	484	2.38	521
ethanol ($\epsilon = 24.3$)	2.74	453 **(425)**	2.46	504 **(542)**	2.30	538
acetonitrile ($\epsilon = 36.3$)	2.74	453 **(422*)**	2.47	502 **(572*)**	2.31	537
DMSO ($\epsilon = 47.2$)	2.71	458 **(422*)**	2.43	511 **(572*)**	2.27	545
water ($\epsilon = 80$)	2.74	453 **(418)**	2.47	502 **(520)**	2.32	537
Trans Isomers						
ethanol ($\epsilon = 24.3$)	2.83	438 **(425)**	2.60	476 **(542)**	2.46	504
water ($\epsilon = 80$)	2.84	437 **(418)**	2.61	474 **(520)**	2.30	538

[a] Shown in bold are the experimental results. By asterisk, we distinguish the wavelengths measured in DMF ($\epsilon = 38.3$).

the experimental band position is 35 nm (0.23 eV). The anionic form of the chromophore observed at neutral and slightly basic pH values absorbs at 520 nm (2.38 eV) in experiments. Theoretically, this band could be assigned to either the cis anion (501 nm or 2.47 eV) or the trans zwitterion (500 nm or 2.48 eV) conformations, giving rise to about 0.1 eV errors. Interestingly, while all TDDFT calculations for molecules in vacuo overestimate excitation energies compared to experimental data, our simulations for the neutral form in aqueous solution result in a value which is underestimated by 0.23 eV. In accord with experimental results, the computed intensities of the neutral form are lower that those of the ionic form. Our calculations also show that the absorption bands in aqueous solution are red-shifted compared to the gas-phase positions by 14−23 nm, depending on the protonation state of the chromophore.

An agreement between our theoretical estimates by using TDDFT and PCM models and experimental results[9] for shifts of the band assigned to the neutral form in various solvents is reasonable. In both studies, the band position is predicted to be slightly sensitive to the solvent. However, an obvious discrepancy is noticed for other solvents, since in calculations the largest shift compared to the value in water ($\epsilon = 80$, $\mu = 1.85$ D) occurs for DMSO ($\epsilon = 47.2$, $\mu = 3.96$ D), but in experiments, it occurs for ethanol ($\epsilon = 24.3$, $\mu = 1.69$ D). For the band associated with the anionic form, the agreement is much worse. Experimentally, the large shifts are observed when moving from 520 nm in water ($\epsilon = 80$) to 542 nm in ethanol and to 572 nm in DMF ($\epsilon = 38.3$, $\mu = 3.82$ D). Our simulations result in the band position lying within 10 nm at most for all considered solvents: water, ethanol, acetonitrile ($\epsilon = 36.6$, $\mu = 3.92$ D), and DMSO ($\epsilon = 47.2$, $\mu = 3.96$ D). The only common observation is that both in theory and in experiment no correlation with either dielectric constant or dipole moment of the solvent molecules is seen.

The authors of ref 9 suggested that the solvent protonic acidity rather than the solvent polarity accounts for observed shifts in band positions in different solvents. In particular, they assumed that "hydrogen bonds between the negatively charged phenolic oxygen and its surrounding solvent shells substantially increase the energy required for excitation".[9] We verified this hypothesis by performing calculations for an extended molecular model, illustrated in Figure 3. We added seven water molecules in order to saturate hydrogen bonds of electronegative atoms and optimized (in solution) geometry parameters of the entire system in the B3LYP/6-31+G(d,p) approximation.

The obtained structures A and B differ by arrangements of the solvent molecules near the imodazole part of the chromophore. Structure A possesses negligibly lower total free energy in solution by 0.02 kcal/mol. In both structures, the phenolic oxygen is involved in the hydrogen bond network with the neighboring solvent molecules; however, the TDDFT [B3LYP/6-311++G(2df,p)] calculations in aqueous solution do not show noticeable changes in excitation energy. Compared to the case of implicit solvation within the continuum model (Table 3: $\Delta E = 2.614$ eV, $\lambda = 474$ nm), a new model results in the following values: $\Delta E = 2.608$ eV and $\lambda = 475$ nm for structure A and $\Delta E = 2.58$ eV and $\lambda = 480$ nm for structure B. Therefore, although the inclusion of explicit solvent molecules in the continuum model may account for slight changes in band position (~5 nm), hydrogen bonding of the chromophore with solvent species can hardly cause a substantial increase of excitation energy as suggested in ref 9.

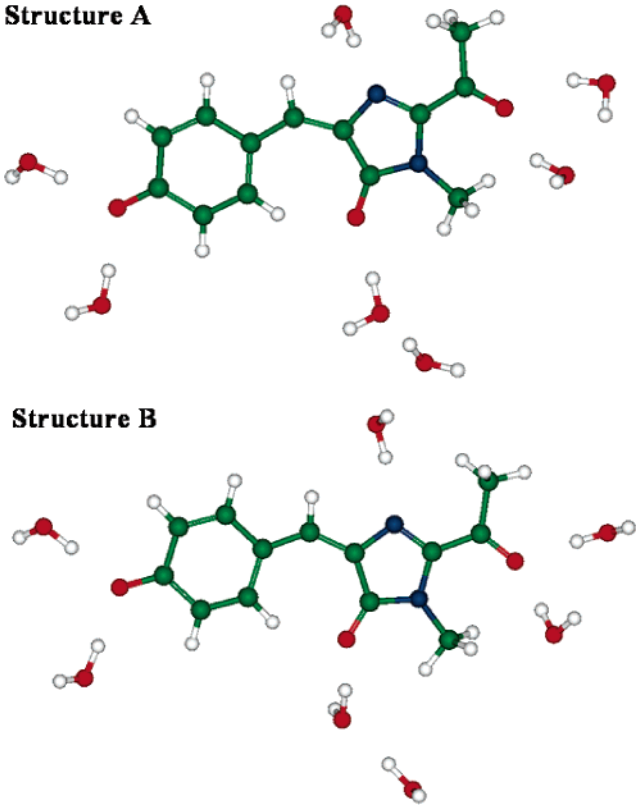


Figure 3. Equilibrium geometry configurations of the trans anion of AHBMI with seven explicit water molecules.

In previous paragraphs, we cited the computed band positions for only cis isomers since they possess lower energies (except for the zwitterionic form). Theoretical calculations of pK_a values of the GFP chromophore[32,33] confirm that all protonation states may occur in solutions. The data obtained in NMR studies of the GFP chromophore in aqueous solution[30] show the barriers for cis−trans transitions for neutral and anionic forms as high as 10 kcal/mol or larger; however, such cis−trans isomerization cannot be excluded as discussed, for instance, in the paper of Xie and Zeng.[27] Although we investigated, in this work, the trans isomers as well, we could not find better agreement with the results of measurements than that presented above for the cis isomers.

We refer the last comment in our discussion to the statement in the paper of Yampolsky et al.[9] about an assignment of the observed weak red fluorescence of the model chromophore in DMF at 603 nm to the same process of absorption−emission as in the native protein asFP595. From simulations described in this paper and elsewhere,[35] this is hard to justify. In protein, the chromophore apparently resides in the trans conformation[6−8] in contrast to the solvent (the observed band at 520 nm in water is uniformly shifted by varying the solvent,[9] and therefore, there are no reasons to assume that in DMSO the chromophore is not in the cis

form). On the other hand, all calculations for gas-phase solutions and protein[35] predict greater wavelengths for cis structures compared to those of the trans species. From both theory[35] and experiment,[8] it follows that the chromophore in asFP595 is excited in the trans conformation and emits in the cis form. Therefore, the observed weak fluorescence in DMF[9] most likely should be assigned to another species.

In conclusion, we report calculations of the spectral properties of the chromophore, AHBMI, from the kindling fluorescent protein asFP595 in vacuo, water, ethanol, acetonitrile, and dimethylsulfoxide in various protonation states in cis and trans conformations by using the TDDFT [B3LYP/6-311++G(2df,p)] and PCM models for geometry parameters optimized for each environment in the B3LYP/6-31+G(d,p) approximation. Despite the simple treatment of the solvent shifts by the dielectric continuum model,[36] the calculation results agree with the majority of conclusions formulated in the experimental studies of this chromophore in water, ethanol, and dimethylformamide.[9] However, some discrepancies with experimental results are underlined with respect to the solvent shifts of the ionic form of the chromophore, as well as with the interpretation of the weak fluorescence in DMF.

Acknowledgment. We thank Dr. A. Savitsky for attracting our attention to this topic and Dr. B. Grigorenko for helpful discussions of the details of this work. We thank the anonymous referees for valuable comments on the original version of the manuscript. This study was partially supported by Grant 04-03-32007 from the Russian Foundation for Basic Research. We thank the staff and administration of the Advanced Biomedical Computing Center for their support of this project. This project has been funded in whole or in part with federal funds from the National Cancer Institute, National Institutes of Health, under Contract NO1-CO-12400. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organization imply endorsement by the U.S. Government.

Supporting Information Available: Cartesian coordinates of the computed structures. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Review: Tsien, R. Y. *Annu. Rev. Biochem.* **1998**, *67*, 509−544.

(2) Review: Zimmer, M. *Chem. Rev.* **2002**, *102*, 759−781.

(3) Review: Schmid, J. A.; Neumeier, H. *ChemBioChem* **2005**, *6*, 1149−1156.

(4) Shaner, N. C.; Campbell, R. E.; Steinbach, P. A.; Giepmans, B. N. G.; Palmer, A. E.; Tsien, R. Y. *Nat. Biotechnol.* **2004**, *22*, 1567−1572.

(5) Lukyanov, K. A.; Fradkov, A. F.; Gurskaya, N. G.; Matz, M. V.; Labas, Y. A.; Savitsky, A. P.; Markelov, M. L.; Zaraisky, A. G.; Zhao, X.; Fang, Y.; Tan, W.; Lukyanov, S. A. *J. Biol. Chem.* **2000**, *275*, 25879−25882.

(6) Quillin, M. L.; Anstrom, D. M.; Shu, X.; O'Leary, S.; Kallio, K.; Chudakov, D. M.; Remington, S. J. *Biochem.* **2005**, *44*, 5774−5787.

(7) Wilmann, P. G.; Petersen, J.; Devenish, R. J.; Prescott, M.; Rossjohn, J. *J. Biol. Chem.* **2005**, *280*, 2401−2404.

(8) Andersen, M.; Wahl, M. C.; Stiel, A. C.; Gräter, F.; Schäfer, L. V.; Trowitzsch, S.; Weber, G.; Eggeling, C.; Grubmüller, H.; Hell, S. W.; Jakobs, S. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 13070−13074.

(9) Yampolsky, I. V.; Remington, S. J.; Martynov, V. I.; Potapov, V. K.; Lukyanov, S.; Lukyanov, K. A. *Biochem.* **2005**, *44*, 5788−5793.

(10) Stratmann, R. E.; Scuseria, G. E.; Frish, M. J. *J. Chem. Phys.* **1998**, *109*, 8218−8224.

(11) Casida, M. E.; Jamorski, C.; Casida, K. C.; Salahub, D. R. *J. Chem. Phys.* **1998**, *108*, 4439.

(12) Tomasi, J.; Mennucci, B.; Cances, E. *THEOCHEM* **1999**, *464*, 211.

(13) Cossi, M.; Barone, V. *J. Chem. Phys.* **2001**, *115*, 4708−4717.

(14) Voityuk, A. A.; Michel-Beyerle, M.-E.; Rösch, N. *Chem. Phys. Lett.* **1997**, *272*, 162−167.

(15) Review: Helms, V. *Curr. Opin. Struct. Biol.* **2002**, *12*, 169−175.

(16) Das, A. K.; Hasegawa, J.-Y.; Miyahara, T.; Ehara, M.; Nakatsuji, H. *J. Comput. Chem.* **2003**, *24*, 1421−1431.

(17) Marques, M. A. L.; Lopez, X.; Varsano, D.; Castro, A.; Rubio, A. *Phys. Rev. Lett.* **2003**, 258101.

(18) Laino, T.; Nifosi, R.; Tozzini, V. *Chem. Phys.* **2004**, *298*, 17−28.

(19) Toniolo, A.; Olsen, S.; Manohar, L.; Martinez, T. J. *Faraday Discuss.* **2004**, *127*, 149−163.

(20) Martin, M. E.; Negri, F.; Olivucci, M. *J. Am. Chem. Soc.* **2004**, *126*, 5452−5464.

(21) Vendrell, O.; Gelabert, R.; Moreno, M.; Lluch, J. M. *Chem. Phys. Lett.* **2004**, *396*, 202−207.

(22) Altoe, P.; Bernardi, F.; Garavelli, M.; Orlandi, G.; Negri, F. *J. Am. Chem. Soc.* **2005**, *127*, 3952−3963.

(23) Sinicropi, A.; Andruniow, T.; Ferre, N.; Basosi, R.; Olivucci, M. *J. Am. Chem. Soc.* **2005**, *127*, 11534−11535.

(24) Lopez, X.; Marques, M. A. L.; Castro, A.; Rubio, A. *J. Am. Chem. Soc.* **2005**, *127*, 12329−12337.

(25) Gross, L. A.; Baird, G. S.; Hoffman, R. C.; Baldridge, K. K.; Tsien, R. Y. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 11990−11995.

(26) Wilmann, P. G.; Petersen, J.; Pettikiriarachchi, A.; Buckle, A. M.; Smith, S. C.; Olsen, S.; Perugini, M. A.; Devenish, R. J.; Prescott, M.; Rossjohn, J. *J. Mol. Biol.* **2005**, *349*, 223−237.

(27) Xie, D. Q.; Zeng, J. *J. Comput. Chem.* **2005**, *26*, 1487.

(28) Nielsen, S. B.; Lapierre, A.; Andersen, J. U.; Pedersen, U. V.; Tomita, S.; Andersen, L. H. *Phys. Rev. Lett.* **2001**, *87*, 228102.

(29) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.;

Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03*, revision B.04; Gaussian, Inc.: Pittsburgh, PA, 2003.

(30) He, X.; Bell, A. F.; Tonge, P. J. *FEBS Lett.* **2003**, *549*, 35−38.

(31) Toniolo, A.; Granucci, G.; Martinez, T. J. *J. Phys. Chem. A* **2003**, *107*, 3822−3830.

(32) Niwa, H.; Inouye, S.; Hirano, T.; Matsuno, T.; Kojima, S.; Kubota, M.; Ohashi, M.; Tsuji, F. I. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 13617−13622.

(33) Yazal, J. E.; Prendergast, F. G.; Shaw, D. E.; Pang, Y.-P. *J. Am. Chem. Soc.* **2000**, *122*, 11411−11415.

(34) Scharnagl, C.; Raupp-Kossmann, R. A. *J. Phys. Chem. B* **2004**, *108*, 477−489.

(35) Grigorenko, B. L.; Nemukhin, A. V.; Savitsky, A. P.; Topol, I. A.; Burt, S. K. To be submitted.

(36) The anonymous reviewer of this paper suggested a helpful analysis of the results of these calculations specifying possible pitfalls of the simple dielectric continuum model, especially, when applied to the charged solute species.

CT050243N

JCTC Journal of Chemical Theory and Computation

A Fast Implementation of Perfect Pairing and Imperfect Pairing Using the Resolution of the Identity Approximation

Alex Sodt, Greg J. O. Beran, Yousung Jung, Brian Austin, and Martin Head-Gordon

Department of Chemistry, University of California, Berkeley, and Chemical Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720-1460

Received September 22, 2005

Abstract: We present an efficient implementation of the perfect pairing and imperfect pairing coupled-cluster methods, as well as their nuclear gradients, using the resolution of the identity approximation to calculate two-electron integrals. The perfect pairing and imperfect pairing equations may be solved rapidly, making integral evaluation the bottleneck step. The method's efficiency is demonstrated for a series of linear alkanes, for which we show significant speed-ups (of approximately a factor of 10) with negligible error. We also apply the imperfect pairing method to a model of a recently synthesized stable singlet biradicaloid based on a planar Ge−N−Ge−N ring, confirming its biradical character, which appears to be remarkably high.

1. Introduction

Resolution of the identity (RI) or density fitting (DF) methods trace their lineage back to early attempts to approximate two-center, four-electron integrals.[1−3] For example, as early as 1939, Sklar used bond-centered auxiliary functions to simplify integral evaluation in an analysis of benzene.[1] In 1966, Harris and Rein approximated two-center function products as sums of one-center function products, determining the auxiliary expansion coefficients by fitting Coulomb integrals, rather than by an overlap criterion.[4] In 1971, Billingsley and Bloor approximated two-center AB products with a linear combination of functions centered on A and B, using what is essentially the procedure we use today: inverting the auxiliary basis Coulomb interaction matrix.[5] In 1969, Newton[6] and, in 1973, Baerands et al.,[7] performed a least-squares fit of the density for their self-consistent field (SCF) calculations, which is now termed[8] the "**S**" approximation, as it effectively minimizes the squared deviation in the overlap (an overlap matrix typically is called an **S** matrix) of the density minus its fit. Also in 1973, Whitten provided theorems bounding the error of least-squares integral fitting, in the Coulomb metric.[9]

In 1979, Dunlap et al. performed a bounded fit of the density for use in $\chi\alpha$. calculations.[10] They minimized the Coulomb self-repulsion of the density minus its fit, a positive semidefinite quantity. They term bounded fits of this nature to be "robust", an adjective we consider to be apt. Dunlap concluded that fitting the electric field generated by electrons is better than fitting the electron density, in the sense that it eliminates first-order error in the fit.[10] This is now termed the "**V**" approximation, where **V** typically denotes an inner product of the Coulomb operator.[8]

Toward applying the RI approximation for general two-electron fitting (rather than fitting a density), which would be useful for correlated wave function theories, Vahtras et al. performed numerical tests of both the **S** and **V** approximations. They found that the **V** approximation reproduced the SCF energy quite accurately, even for modestly sized auxiliary basis sets.[8]

In 1995, Eichkorn et al. produced an auxiliary basis set which reproduced the **J** matrix to reasonable accuracy, with an auxiliary basis set of approximately 3 times the size of the orbital basis set.[11] Optimized basis sets for use with MP2 followed,[12,13] which were able to reproduce results within a few μH per atom.

One expects such success from this density fitting procedure because, while the linear combination of atomic orbitals (LCAO) basis set (denoted by greek indices) used might not be very linearly dependent, the product space (termed $\mu\nu$ products) almost certainly will be.[14,15] This two-center

10.1021/ct050239b CCC: $33.50
Published on Web 01/11/2006

product space is then very amenable to an expansion in a much smaller basis (or perhaps by simply eliminating the linearly dependent portion of the space outright[14−16]).

The computational advantage of the RI approximation is not only that it reduces the four-center integrals to a composite of three-center ones but also that it separates the two-electron integral into a contraction over two one-electron expansions, which can be transformed to a molecular orbital representation independently. It is, thus, an indispensable tool for methods limited both by integral computation (such as Hartree−Fock theory[17]) and by temporary storage for the atomic orbital (AO) to molecular orbital (MO) transformation (such as MP2).[18] The RI method has some similarities with the pseudospectral method, which has also been implemented for perfect pairing[19] (PP). Relative to the pseudospectral approach, RI methods have the advantage of being completely smooth and, thus, so too are the potential energy surfaces that result.

We apply the RI approximation to perhaps the most basic of correlated wave function methods, PP[20,21] and imperfect pairing (IP), which are described in more detail elsewhere.[22−24] Both of these methods can be viewed as approximations to valence-optimized doubles (VOD),[25] itself an approximation to a complete valence space treatment, complete active space−self-consistent field (CASSCF). VOD represents a triumph in that it scales to the sixth order (as opposed to exponential scaling, like CASSCF), yet even sixth-order scaling practically imposes a hard wall beyond which we cannot apply the method.

In coupled-cluster theory, the ground-state trial wave function is written as an exponentiated excitation operator acting on a reference state, $|0\rangle$:

$$|\Psi\rangle = \exp(\hat{T})|\,0\rangle \tag{1}$$

where, for VOD, the $\hat{T}$ operator is

$$\hat{T}_{\mathrm{VOD}} = \sum_{ijk^*l^*} t_{ij}^{k^*l^*} a_{k^*}^{\dagger} a_{l^*}^{\dagger} a_j a_i \tag{2}$$

The orbital occupation creation and annihilation operators ($a^{\dagger}$ and a, respectively) are weighted by so-called t amplitudes, t.

PP truncates the excitation operator to include what would presumably be the most important double excitations. Each active α electron is paired with exactly one β electron, and they are simultaneously correlated with exactly one pair of virtual orbitals. In this way, pairs of electrons are correlated independently of each other. The form of the PP excitation operator is

$$\hat{T}_{\mathrm{PP}} = \sum_i t_{i\bar{i}}^{i^*\bar{i}^*} a_{i^*}^{\dagger} a_{\bar{i}^*}^{\dagger} a_{\bar{i}} a_i \tag{3}$$

This operator should perform well for breaking isolated bonds; as occupied and virtual orbitals become nearly degenerate, the PP wave function will be able to contain a mixture of the two.

IP truncates the coupled-cluster doubles (CCD) excitation operator such that correlation is provided between the most important pairs. IP allows one electron to be excited from each of two pairs simultaneously. The form of the IP excitation operator for a system with an even number of electrons is

$$\hat{T}_{\mathrm{IP}} = \hat{T}_{\mathrm{PP}} + \sum_{i\neq j}\left(t_{i\bar{j}}^{i^*\bar{j}^*} a_{i^*}^{\dagger} a_{\bar{j}^*}^{\dagger} a_{\bar{j}} a_i + t_{i\bar{j}}^{i^*\bar{i}^*} a_{j^*}^{\dagger} a_{\bar{i}^*}^{\dagger} a_{\bar{j}} a_i + \frac{1}{2} t_{ij}^{i^*j^*} a_{i^*}^{\dagger} a_{j^*}^{\dagger} a_j a_i + \frac{1}{2} t_{\bar{i}\bar{j}}^{\bar{j}\bar{i}^*} a_{\bar{j}^*}^{\dagger} a_{\bar{j}^*}^{\dagger} a_{\bar{j}} a_{\bar{i}} \right) \tag{4}$$

This operator retains the desirable properties of PP, yet will also correlate important open-shell configurations. It also provides interpair correlation, which is physically important in cases such as multiple bonding. Also, for systems that do not resemble a group of localized electron pairs, it will provide a more physically consistent description than PP. Van Voorhis and Head-Gordon explained this illustrative example in their development of IP: Benzene's π electrons are completely delocalized across the six-membered ring. Instead of selecting delocalized correlating orbitals that reflect this, PP instead localizes the electrons to maximize the pair correlation. The extra flexibility granted to IP remedies this problem to a large extent, but there is still spurious symmetry breaking as the method still places too much emphasis on the most important pair excitations.[23]

The coupled-cluster equations are solved iteratively, in a process which, for a full doubles treatment, scales to the sixth power of system size (o^2v^4). In contrast, the effect of the limited number of PP and IP amplitudes is to make two-electron integral construction (scaling to the fourth power of system size) more expensive than actually solving the equations (for IP, this scales to the third power of the number of pairs, while in PP, the amplitudes are independent of each other). While the integrals are limited in number, they must still be formed from many integrals over atomic orbitals. It is just such a case for which the RI approximation should prove most successful; it provides a total reduction in the number of base integrals computed, and it may transform each electron of the two-electron integral independently.

The PP and IP orbitals are optimized such that they provide the lowest total energy (this also serves to define a unique pairing scheme). This is accomplished by forming the gradient of the energy with respect to orbital rotation and then performing standard search techniques to minimize the energy. Taking this orbital gradient requires an expanded set of Coulomb integrals, all of which can be obtained from the following half-transformed integrals:[19,22]

$$\begin{gathered}
\mathbf{J}_{\mu\nu}^{ii} = (ii|\mu\nu) \\
\mathbf{J}_{\mu\nu}^{ii^*} = (ii^*|\mu\nu) \\
\mathbf{J}_{\mu\nu}^{i^*i^*} = (i^*i^*|\mu\nu) \\
\mathbf{K}_{\mu\nu}^{ii} = (i\mu|i\nu) \\
\mathbf{K}_{\mu\nu}^{ii^*} = (i\mu|i^*\nu) \\
\mathbf{K}_{\mu\nu}^{i^*i^*} = (i^*\mu|i^*\nu)
\end{gathered} \tag{5}$$

2. Algorithm

In developing an RI approximation for computing these intermediates, we first define the auxiliary basis expansion of a single function product, $|\mu\nu)$. We minimize the self-interaction of the product minus its fit:

$$(\mu\nu - \overline{\mu\nu}|\mu\nu - \overline{\mu\nu}) \tag{6}$$

This leads to the following (optimal[26]) expression for the four-center two-electron integrals in terms of a set of three-center quantities, the **B** tensors, which are given both in terms of atomic orbitals (Greek letters) and in auxiliary basis functions (L, M...):

$$(\mu\nu|\lambda\sigma) \approx \sum_L \mathbf{B}^L_{\mu\nu}\mathbf{B}^L_{\lambda\sigma} = \sum_{LMN}(\mu\nu|L)(L|M)^{-1/2}(M|N)^{-1/2}(N|\lambda\sigma) \tag{7}$$

where a **B** tensor[12] is defined to be

$$\mathbf{B}^L_{\mu\nu} = \sum_K (\mu\nu|K)(K|L)^{-1/2} \tag{8}$$

and where $\mu\nu$ need not necessarily be atomic orbitals but could, in fact, be transformed into molecular orbitals. By expanding two-center functions in terms of one-center auxiliary functions, explicit four-center integrals are never needed. The number of two-center function products, which we call NFP, formally scales linearly with system size, due to the fact that Gaussian AOs have limited spatial extent, and therefore, the product of two well-separated AOs will be negligible. Thus, the number of required two-electron integrals will be reduced, but will still scale quadratically with system size. The task is, thus, to form the requisite **B** tensors most efficiently and, then, to transform them into the **J** and **K** matrices. The following is an outline of the RI algorithm, with the scaling of the step indicated in parentheses.

1a*. Form: $(L|M)^{-1/2}$ (X^3)

2a*. Form: $(\mu\nu|M)$ (NFP X)

3a*. Contract: $\mathbf{B}^L_{\mu\nu} = \sum_M (\mu\nu|M)(M|L)^{-1/2}$ (NFP X^2)

4a. Contract: $\mathbf{B}^L_{\mu[i,i^*]} = \sum_\nu \mathbf{B}^L_{\mu\nu}C_{\nu[i,i^*]}$ (NFP $X\,o$)

5a. Contract: $\mathbf{B}^L_{[ii,ii^*,i^*i^*]} = \sum_\mu \mathbf{B}^L_{\mu[i,i^*]}C_{\mu[i,i^*]}$ ($X\,N\,o$)

6a. Contract: $\mathbf{K}^{[ii,ii^*,i^*i^*]}_{\mu\nu} = \sum_L \mathbf{B}^L_{\mu[i,i^*]}\mathbf{B}^L_{\nu[i,i^*]}$ ($X\,N^2\,o$)

7a. Contract: $\mathbf{J}^{[ii,ii^*,i^*i^*]}_{\mu\nu} = \sum_L \mathbf{B}^L_{\mu\nu}\mathbf{B}^L_{[ii,ii^*,i^*i^*]}$ (NFP $X\,o$)

Steps marked with an asterisk need only be computed once per calculation, otherwise the step must be done each time MOs are updated. Comma-separated indices in square brackets are independent, and their contributions to the total cost of the step are, therefore, mutually additive. Contrast this with the alternative algorithm for creating the **J** and **K** matrices using four-center AO integrals without the RI approximation:

1b. Form: $(\mu\nu|\lambda\sigma)$ (NFP^2)

2b. Contract: $\mathbf{K}^{[ii,ii^*,i^*i^*]}_{\mu\nu} = \sum_{\lambda\sigma}(\mu\lambda|\nu\sigma)C_{\lambda[i,i^*]}C_{\sigma[i,i^*]}$ $(\mathrm{NFP}^2\cdot o)$

3b. Contract: $\mathbf{J}^{[ii,ii^*,i^*i^*]}_{\mu\nu} = \sum_{\lambda\sigma}(\mu\nu|\lambda\sigma)C_{\lambda[i,i^*]}C_{\sigma[i,i^*]}$ $(\mathrm{NFP}^2\cdot o)$

Our RI algorithm formally scales with the fourth power of system size (step 6a), while without the RI approximation, the algorithm scales only with the third power of system size (steps 2b and 3b). However, for systems of a size for which either algorithm might be feasible, few function products can be neglected. NFP is, thus, comparable to N^2, yielding effectively fifth-order scaling without the RI approximation. In fact, step 4a of the RI algorithm is the dominant step for small to modestly large systems, both because of the size of NFP and because step 6a is implemented as a matrix multiply, employing optimized standard routines. Step 4a cannot be simply computed as a matrix multiply without breaking the inherent sparsity of the function product. Each step of the RI algorithm may be broken up such that it can be computed using limited memory that scales quadratically with system size. Mass storage requirements scale cubically with system size. Contrast these modest requirements with the coupled-cluster doubles methods that do not employ local truncations, which have large (quartic) disk requirements and large (sixth-order) computational requirements.

We implemented these methods into a development version of the quantum chemistry package, Q-CHEM.[27]

A comparison of the performance of RI−PP, RI−IP, PP, and IP is shown in Table 1. It is evident that the resolution of the identity approximation reduces computation time by approximately one order of magnitude. Also, because the bottleneck step of both PP and IP is the construction of the same integral intermediates, they are both accelerated to the same degree, and indeed, there is little difference in the computational requirements between the methods. The effect of the RI approximation is more pronounced in the larger basis set because the auxiliary basis size does not increase

Table 1. Total CPU Times Comparing Resolution of the Identity (RI) Algorithms for the PP and IP Methods, against a Non-RI PP Algorithm on Linear Alkanes[a]

chain length (basis)	RI−PP CPU (s)	RI−IP CPU (s)	PP CPU (s)	IP CPU (s)
2 (cc-pVDZ)	8.0	8.8	25.0	32.5
2 (cc-pVTZ)	321.3	342.1	1267.6	1373.2
4 (cc-pVDZ)	54.1	63.2	281.1	344.5
4 (cc-pVTZ)	2644.7	2809.7	20688.1	22216.2
6 (cc-pVDZ)	170.4	198.7	1120.1	1564.6
8 (cc-pVDZ)	321.1	402.6	3061.8	4321.6

[a] Calculations were performed using a single 2.3 GHz IBM 970fx processor in an Apple XServe. The basis set used is cc-pVDZ with its RI-MP2 fitting basis.[28] Each calculation required between 11 and 13 iterations.

proportionally with the size of the AO basis. Larger basis sets should be relatively easier to fit, because the linear dependence of the product space ($\mu\nu$) will be greater.

The RI approximation introduces very little error to the PP and IP methods. As an example, for the molecules of the G2 set using the SV[29] basis, the RI approximation (using the algorithm described above) introduces a root-mean-square (RMS) error of 29 μH to the PP energy, or 8 μH per atom. The accuracy of the method is fairly uniform over the set of molecules, the greatest error being 31 μH per atom. For the closed-shell subset of the G2 set, RI−IP gives an RMS error of 64 μH, or 13 μH per atom. The largest error for RI−IP is 71 μH per atom.

3. Nuclear Gradient

In PP and IP, the nuclear gradient is formed from a similar set of intermediate integrals as the orbital gradient:

$$\mathbf{J}^{ii(x)}_{\mu\nu} = (ii|\mu^{(x)}\nu) + (ii|\mu\nu^{(x)})$$
$$\mathbf{J}^{ii^*(x)}_{\mu\nu} = (ii^*|\mu^{(x)}\nu) + (ii^*|\mu\nu^{(x)})$$
$$\mathbf{J}^{i^*i^*(x)}_{\mu\nu} = (i^*i^*|\mu^{(x)}\nu) + (i^*i^*|\mu\nu^{(x)})$$
$$\mathbf{K}^{ii(x)}_{\mu\nu} = (i\mu^{(x)}|i\nu) + (i\mu|i\nu^{(x)})$$
$$\mathbf{K}^{ii^*(x)}_{\mu\nu} = (i\mu^{(x)}|i^*\nu) + (i\mu|i^*\nu^{(x)})$$
$$\mathbf{K}^{i^*i^*(x)}_{\mu\nu} = (i^*\mu^{(x)}|i^*\nu) + (i^*\mu|i^*\nu^{(x)}) \quad (9)$$

The RI algorithm for constructing these integrals follows directly from the RI algorithm above, but acting on six sets of integrals, with *x*, *y*, and *z* derivatives with respect to the center of both μ and ν. However, one must also consider the effect of nuclear displacement on the auxiliary basis function centers. Here is the nuclear gradient of the energy:

$$E^{(x)} = E_0^{(x)} + \sum_i (f_i^{i(x)}\gamma_i^i + f_{i^*}^{i^*(x)}\gamma_{i^*}^{i^*}) + \sum_{pqrs}\Gamma_{pqrs}(pq|rs)^{(x)}_{RI} \quad (10)$$

where the two-electron integral contracted with the effective two-particle density matrix, Γ, is fit using the RI approximation, and the sum over orbitals (*p*, *q*, *r*, and *s*) is restricted to those relevant to the pairing method used.

Derivatives with respect to regular basis function centers are included in the **J** and **K** terms described above. We begin by stating this useful identity for computing derivatives with respect to auxiliary basis centers:

$$\frac{\partial}{\partial x}(K|L)^{-1} = -\sum_{M,N}(K|M)^{-1}\left[\frac{\partial}{\partial x}(M|N)\right](N|L)^{-1} \quad (11)$$

We use the following intermediates for efficiently constructing the gradient:

$$\tilde{\Gamma}^L_{pq} = \sum_{rs,L}(L|M)^{-1}(M|rs)\Gamma_{pqrs} \quad (12)$$

The result is then added to the gradient (the subscript RI indicates that this is only the contribution from the gradient with respect to auxiliary basis center):

$$\tilde{\Gamma}^{KL} = -\sum_{pq}(pq|M)(M|K)^{-1}\tilde{\Gamma}^K_{ab} \quad (13)$$

$$E^{(x)}_{RI} = 2\sum_{pq,K}(pq|K^{(x)})\tilde{\Gamma}^K_{ab} + 2\sum_{KL}(K^{(x)}|L)\tilde{\Gamma}^{KL} \quad (14)$$

For PP and IP, the effective two-particle density matrix is sparse. For PP, there is a linear number of nonzero terms: Γ_{iiii}, $\Gamma_{i^*i^*i^*i^*}$, $\Gamma_{iii^*i^*}$, $\Gamma_{ii^*i^*i}$, $\Gamma_{ii^*ii^*}$, and $\Gamma_{i^*i^*ii}$. For IP, there is a quadratic number of nonzero terms: Γ_{iijj}, Γ_{ijji}, $\Gamma_{iij^*j^*}$, $\Gamma_{ij^*j^*i}$, $\Gamma_{i^*i^*j^*j^*}$, $\Gamma_{i^*j^*j^*i^*}$, $\Gamma_{ii^*jj^*}$, $\Gamma_{ij^*ji^*}$, and $\Gamma_{i^*j^*ji}$. This significantly reduces both the storage and computational requirements of the above intermediates.

Here is a step-by-step description of the RI algorithm for the gradient with respect to auxiliary basis centers, with PP and then IP cost scalings indicated in parentheses where they differ (and assuming three-center integrals are left over from the PP/IP calculation):

1c. Compute: $(K^{(x)}|L)$ (X^2)

2c. Compute: $(K|L)^{-1}$ (X^3)

3c. Transform: $A^K_{\mu\nu} = \sum_L (K|L)^{-1}(L|\mu\nu)$ (NFP X^2)

4c. Transform: $A^K_{\mu q} = \sum_\nu A^K_{\mu\nu}C_{\nu q}$ (NFP $X\,o$)

5c. Transform: $A^K_{pq} = \sum_\mu A^K_{\mu q}C_{\mu p}$ ($N\,X\,o$, $N\,X\,o^2$)

6c. Contract: $\tilde{\Gamma}^K_{pq} = \sum_{rs}\Gamma_{pqrs}A^K_{rs}$ ($X\,o$, $X\,o^2$)

7c. Contract: $\tilde{\Gamma}^{KL} = \sum_{pq}A^K_{pq}\tilde{\Gamma}^L_{pq}$ ($X^2\,o$, $X^2\,o^2$)

8c. Compute: $(\mu\nu|K^{(x)})$ (NFP X)

9c. Transform: $(\mu q|K^{(x)}) = \sum_\nu (\mu\nu|K^{(x)})C_{\nu q}$ (NFP $X\,o$)

10c. Transform: $(pq|K^{(x)}) = \sum_\mu (\mu q|K^{(x)})C_{\mu p}$ ($N\,X\,o$, $N\,X\,o^2$)

11c. Increment/Contract: $\nabla E \leftarrow \sum_{pq,K}(pq|K^{(x)})\tilde{\Gamma}^K_{pq}$ ($X\,o$, $X\,o^2$)

12c. Increment/Contract: $\nabla E \leftarrow \sum_{KL}(K^{(x)}|L)\tilde{\Gamma}^{KL}$ (X^2)

Like for the RI algorithm for the PP/IP energy, the most expensive steps are transformation steps such as 4c, 5c, 9c, and 10c, for modestly large systems. Step 7c is the only new type of fourth-order step introduced, and this only for IP. In total, the calculation of the gradient is costlier than a single iteration of PP/IP, because of the prefactor introduced by transformations for each derivative component. Figure 1 shows a comparison of the required gradient and energy CPU

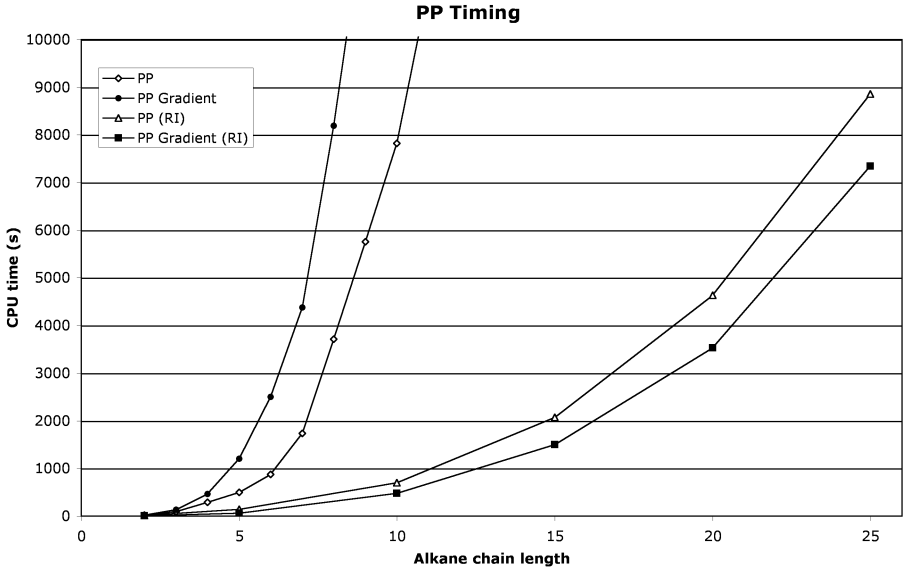


Figure 1. Total CPU time for a series of linear alkanes. Calculations were run on one 2.3 GHz Xserve G5 processor, with 8 GB of RAM, using a cc-pVDZ basis set and its corresponding RIMP2 fitting basis.[28]

time for a series of linear alkanes using the restricted perfect pairing method.

4. An Application of Imperfect Pairing: Prediction of Diradical Character

IP should provide an adequate description of strongly correlated systems, such as open-shell singlet diradicaloids.[30] Species with diradical character are important in that they may be intermediates in chemical reactions. It is also imaginable that they would be useful for their unique properties, in that they would have two weakly coupled unpaired electrons. One desires to apply a high-level correlated treatment to such problems, but stable diradicaloids are often only stable because of steric substituent effects, making calculations on all but the smallest model systems too costly for coupled-cluster doubles theory. As Kohn−Sham DFT,[31] through its representation as a single electron configuration, does not provide for correlated fractional occupation of orbitals except through spin-symmetry breaking, it is of limited use in diagnosing diradical character. As IP is based on coupled-cluster theory, its t amplitudes provide quantitative indications of partial virtual orbital occupation. Furthermore, our efficient implementation of IP has enabled us to explore a broader range of such molecules than before.

We applied the IP method to a model of a stable diradicaloid compound synthesized by Cui et al.[32] In forming the model, we kept a modestly sized portion of a very bulky substituent. The model molecule is shown in Figure 2. It is of a class of diradicaloids with a strained four-member ring, in which two diagonally opposed atoms (in this case, germaniums) have a partially filled valence. The calculation was performed with the SVP[29] basis, and with Ahlrichs' corresponding fitting basis.[12] The HOMO and LUMO calculated with IP are shown in Figure 2. The HOMO and LUMO appear to be out-of-phase nonbonding orbitals, consistent with Cui et al.'s DFT calculations, and also with a simple interpretation of germanium's valence. We found that the HOMO had a fractional occupation of 78.9% and

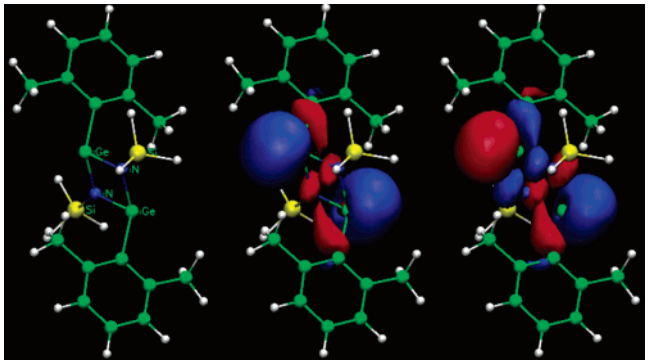


Figure 2. Left to right are the model molecule, the HOMO, and the LUMO.

that the LUMO had 21.1%. Considering the out-of-phase nature of the HOMO and LUMO, and what we consider a good definition of the measure of diradical character,[33] this indicates extremely high diradical character, as far as stable singlet diradicals go. If one imagines that a pure diradical would have 50% HOMO occupation and 50% LUMO occupation, consistent with an H_2 molecule with a stretched bond, 21.1% LUMO occupation means 42.2% diradical character. Although the pool of synthesized stable singlet diradicaloids available for comparison at this point in time is small, we would consider 42.2% diradical character to be quite remarkable. As a comparison, consider the very much reactive Si(100) surface. The cleaved surface rearranges to form Si dimers whose bonds are intermediate between a single and a double bond and which have about 35% diradical character,[34] less than that of Cui et al.'s stable singlet diradicaloid.

We must reiterate that we did not run IP on the full synthesized molecule and that the protective groups could have an important stabilizing effect and could increase the HOMO−LUMO gap, decreasing the diradical character. We will present a more thorough study of this stable singlet diradicaloid in a future work. Still, the IP method here serves to reaffirm the conclusion the authors made (that the

molecule was indeed a singlet diradicaloid), the evidence for which was a strained DFT geometry, spectroscopic data, and its reactivity. Here, IP adds an important indication of diradical character.

5. Conclusion

We have presented an efficient implementation of an RI algorithm for calculating the PP and IP energy, as well as the nuclear gradient for both PP and IP. The RI approximation eliminates the need for four-center two-electron integrals by replacing two-center atomic orbital function products with a sum of one-center auxiliary basis functions. This creates separability between the two electrons of the two-electron integral, allowing one to transform the two coordinates of the integral from an AO representation to an MO representation independently. While this introduces fourth-order scaling due to the RI contraction step, in the regime where PP and IP are feasible, it significantly reduces computational cost by about a factor of 10, while introducing error that is not likely to impact quantitative results. The speed of the method is then very much comparable to SCF theory, at least until linear scaling approximations for Fock matrix construction become useful.

As an example of a calculation readily feasible for PP and IP, yet far more costly for a full doubles treatment, such as CCD, we chose a model of a recently synthesized stable singlet biradical. Our correlated treatment of the system showed significant biradical character for the molecule, a result that DFT can only predict by inference through calculated structural properties.

Acknowledgment. This work was partly supported by the Department of Energy, Office of Basic Energy Sciences, SciDAC Computational Chemistry Program (Grant DE-FG0201ER403301), with additional support from subcontracts from National Institutes of Health Small Business Innovation Research Grants to Q-Chem Inc. M.H.-G. is a part-owner of Q-Chem Inc.

References

(1) Sklar, A. L. *J. Chem. Phys.* **1939**, *7*, 984−993.

(2) Mulliken, R. *J. Chem. Phys. Phys. Chim. Biol.* **1949**, *46*, 497.

(3) Løwdin, P. *J. Chem. Phys.* **1953**, *21*, 374−375.

(4) Harris, F. E.; Rein, R. *Theor. Chem. Acc.* **1966**, *6*, 73−82.

(5) Billingsley, F. P.; Bloor, J. E. *J. Chem. Phys.* **1971**, *55*, 5178−5190.

(6) Newton, M. D. *J. Chem. Phys.* **1969**, *51*, 3917−3926.

(7) Baerands, E. J.; Ellis, D. E.; Ros, P. *Chem. Phys.* **1973**, *2*, 41−51.

(8) Vahtras, O.; Almløf, J.; Feyereisen, M. W. *Chem. Phys. Lett.* **1993**, *213*, 514−518.

(9) Whitten, J. L. *J. Chem. Phys.* **1973**, *58*, 4496−4501.

(10) Dunlap, B. I.; Connolly, J. W. D.; Sabin, J. R. *J. Chem. Phys.* **1979**, *71*, 3396−3402.

(11) Eichkorn, K.; Treutler, O.; Øhm, H.; Häser, M.; Ahlrichs, R. *Chem. Phys. Lett.* **1995**, *240*, 283−289.

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

(13) Bernholdt, D. E.; Harrison, R. J. *J. Chem. Phys.* **1998**, *109*, 1593−1600.

(14) Beebe, N. H. F.; Linderberg, J. *Int. J. Quantum Chem.* **1977**, *12*, 683−705.

(15) O'Neal, D. W.; Simons, J. *Int. J. Quantum Chem.* **1989**, *36*, 673−688.

(16) Ten-no, S.; Iwata, S. *J. Chem. Phys.* **1996**, *105*, 3604−3611.

(17) Weigend, F. *Phys. Chem. Chem. Phys.* **2002**, *4*, 4285−4291.

(18) Bernholdt, D. E.; Harrison, R. J. *Chem. Phys. Lett.* **1996**, *250*, 477−484.

(19) Langlois, J.; Muller, R. P.; Coley, T. R.; Goddard, W. A., III; Ringnalda, M. W.; Won, Y.; Friesner, R. A. *J. Chem. Phys.* **1990**, *92*, 7488−7497.

(20) Cullen, J. *Chem. Phys.* **1996**, *202*, 217−229.

(21) Bobrowicz, F. W.; Goddard, W. A., III. *Modern Theoretical Chemistry: Methods of Electronic Structure Theory*; Schaefer, H. F., III, Ed.; Plenum: New York, 1977; Vol. 3, pp 79−126.

(22) Van Voorhis, T.; Head-Gordon, M. *J. Chem. Phys.* **2002**, *117*, 9190−9201.

(23) Van Voorhis, T.; Head-Gordon, M. *Chem. Phys. Lett.* **2000**, *317*, 575−580.

(24) Beran, G. J. O.; Austin, B.; Sodt, A.; Head-Gordon, M. *J. Phys. Chem. A* **2005**, *109*, 9183−9192.

(25) Krylov, A. I.; Sherrill, C. D.; Byrd, E. F. C.; Head-Gordon, M. *J. Chem. Phys.* **1998**, *109*, 10669−10678.

(26) Dunlap, B. I. *THEOCHEM* **2000**, *529*, 37−40.

(27) Kong, J. *J. Comput. Chem.* **2000**, *21*, 1532−1548.

(28) Weigend, F.; Køhn, A.; Hättig, C. *J. Chem. Phys.* **2002**, *116*, 3175−3183.

(29) Schäfer, A.; Horn, H.; Ahlrichs, R. *J. Chem. Phys.* **1992**, *97*, 2571−2577.

(30) Salem, L.; Rowland, C. *Angew. Chem., Int. Ed.* **1972**, *11*, 92−111.

(31) Kohn, W.; Becke, A. D.; Parr, R. G. *J. Phys. Chem.* **1996**, *100*, 12974−12980.

(32) Cui, C.; Brynda, M.; Olmstead, M. M.; Power, P. P. *J. Am. Chem. Soc.* **2004**, *126*, 6510−6511.

(33) Jung, Y.; Head-Gordon, M. *ChemPhysChem* **2003**, *4*, 522−525.

(34) Jung, Y.; Akinaga, Y.; Jordan, K. D.; Gordon, M. S. *Theor. Chem. Acc.* **2003**, *109*, 268−273.

CT050239B

JCTC Journal of Chemical Theory and Computation

Evaluation of Two-Center, Two-Electron Integrals

Alejandro Ferrón and Pablo Serra*

Facultad de Matemática, Astronomía y Física, Universidad Nacional de Córdoba, Ciudad Universitaria, 5000 Córdoba, Argentina

Received October 26, 2005

Abstract: We present a new analytic treatment of two-electron integrals over two-center integrals including correlation (interelectronic distance) explicitly in the wave function. All the integrals needed for the evaluation of the matrix elements of any diatomic two-electron molecule are obtained as analytic recursion expressions. As an application of this method in molecular physics, we calculate the value of the ground-state energy and equilibrium internuclear distance of the hydrogen molecule in the Born−Oppenheimer approximation.

1. Introduction

Two-center, two-electron systems are a subject of great interest in molecular physics. In particular, several molecules may be described as such a system. Computation of the Born−Oppenheimer ground-state energy of the hydrogen molecule was the subject of progressively more accurate variational calculations.[1−5] Two-electron addition to closed-shell neutral polar molecules may also be described as a two-center, two-electron system. The binding of two electrons to a fixed finite dipole has not been resolved. In recent years, there has been increasing interest in the study of the possible existence of such dipole-bound dianions.[6−9] The study of this kind of weakly bound states represents an interesting field of research. For these states, the energy is nonanalytical as a function of the dipolar moment, and a bound state could not exist at the threshold energy; therefore, they might be good candidates to be halo states.[10] Technical problems appear when standard approximations such as perturbation theory, nonlinear variational calculations, or the Rayleigh−Ritz method are used to study weakly bound states.[11] Recently, a finite size-scaling theory for the study of near-threshold properties in quantum few-body problems has been developed.[12] The method was successfully applied to one electron attached to dipole and quadrupole potentials.[13,14] An accurate expansion of the ground-state wave function in a (truncated) complete basis-set is necessary in order to apply finite size-scaling methods to two-center, two-electron systems.

James and Coolidge[1] were the first that made ab initio calculations for two-center, two-electron systems using correlated functions. These functions include the interelectronic coordinate explicitly. They extend the method used by Hylleraas for the helium atom[15] to the hydrogen molecule. After this, many authors used the James−Coolidge or modified James−Coolidge expansions for the calculation of different properties of diatomic two-electron systems.[2,4,5]

Even if the results obtained using James−Coolidge expansion are very accurate, this method has some difficulties. The inclusion of the interelectronic coordinate (as powers) in the wave function generates very complicated two-electron, two-center integrals. Kolos et al.[2,4] solved these integrals keeping powers of the interelectronic distance up to order three and obtained very accurate values for the ground-state energy of the hydrogen molecule.[4] In ref 16, Kolos and Roothaan present an interesting treatment of these integrals. They solve fully analytically the case of even powers of the interelectronic distance. The case of odd powers was partially solved and completed with numerical integration.

The aim of this paper is to report a new method for the analytical evaluation of the two-electron, two-center James−Coolidge integrals without limitation in the power of the correlation coordinate.

This paper is organized as follows. In section 2, we develop our method for the evaluation of two-center, two-electron integrals and we express these integrals as analytical recurrence relations. Technical aspects are discussed and numerical evaluations are presented in section 3. In section 4, we apply the results obtained in previous sections to

* Corresponding author e-mail: serra@famaf.unc.edu.ar; homepage: http://tero.fis.uncor.edu/∼serra/.

10.1021/ct0502662 CCC: $33.50
Published on Web 01/25/2006

evaluate the ground-state energy of the hydrogen molecule. Finally, our conclusions are given in section 5.

2. Two-Electron Integrals over Two-Center Orbitals

The basic integrals that appear in a two-center, two-electron James−Coolidge ground-state expansion of any diatomic two-electron system are of the form[2]

$$I^m_{pqrs} = \left(\frac{2}{R}\right)^6 \int \mathrm{d}^3x_1\, \mathrm{d}^3x_2\, \frac{\mathrm{e}^{-(\alpha\xi_1+\beta\xi_2)}}{\xi_1^2-\eta_1^2}\xi_1^p\eta_1^q\xi_2^r\eta_2^s r_{12}^m \qquad (1)$$

where the integral is expressed in usual prolate spheroidal coordinates (ξ, η, and ϕ)[17]. ϕ is the azimuthal angle, $\xi = (r_a + r_b)/R$, $\eta = (r_a - r_b)/R$, r_a and r_b are the distances to the centers, R represents the distance between centers, r_{12} is the interelectronic distance, and α and β are variational parameters. Powers are integer numbers with $p, q, r, s \geq 0$ and $m \geq -1$.

Introducing the auxiliary integral

$$g_{pqrs}(k) = \left(\frac{2}{R}\right)^6 \int \mathrm{d}^3x_1\, \mathrm{d}^3x_2\, \frac{\mathrm{e}^{-(\alpha\xi_1+\beta\xi_2)}}{\xi_1^2-\eta_1^2}\xi_1^p\eta_1^q\xi_2^r\eta_2^s \frac{\mathrm{e}^{ikr_{12}}}{r_{12}} \qquad (2)$$

eq 1 may be expressed as

$$I^m_{pqrs} = \frac{1}{i^{m+1}} \frac{\partial^{m+1}}{\partial k^{m+1}} g_{pqrs}(k)\Big|_{k=0} \qquad (3)$$

Then, the problem is reduced to solve the integral in eq 2. For this purpose, we use the expansion of the Green function for the Helmholtz operator in prolate spheroidal coordinates:[17]

$$\frac{e^{ikr_{12}}}{r_{12}} = 4\pi ik \sum_{m,l} \frac{(2l+1)(l-m)!}{4\pi(l+m)!} S^{(1)}_{ml}\left(\frac{k}{2},\eta_1\right) S^{(1)}_{ml}\left(\frac{k}{2},\eta_2\right) \mathrm{e}^{im(\phi_1-\phi_2)}\, je_{ml}\left(\frac{k}{2},\xi_<\right) he_{ml}\left(\frac{k}{2},\xi_>\right) \qquad (4)$$

The functions present in this expansion are the *spheroidal wave functions*.[18,19] Replacing eq 4 in eq 2 and integrating over ϕ_1 and ϕ_2, we obtain

$$g_{pqrs}(k) = 4\pi^2 ik \sum_l (2l+1) \int \mathrm{d}^2x_1\, \mathrm{d}^2x_2\, S^{(1)}_{0l}\left(\frac{k}{2},\eta_1\right) S^{(1)}_{0l}\left(\frac{k}{2},\eta_2\right) R^{(1)}_{0l}\left(\frac{k}{2},\xi_<\right)\left[R^{(1)}_{0l}\left(\frac{k}{2},\xi_>\right) + iR^{(2)}_{0l}\left(\frac{k}{2},\xi_>\right)\right] f_{pqrs}(\xi_1,\eta_1,\xi_2,\eta_2) \qquad (5)$$

where

$$f_{pqrs}(\xi_1,\eta_1,\xi_2,\eta_2) = \frac{\mathrm{e}^{-(\alpha\xi_1+\beta\xi_2)}\xi_1^p\,\eta_1^q\,\xi_2^r\,\eta_2^s}{\xi_1^2-\eta_1^2} \qquad (6)$$

$$je_{0l}(c,\xi) = R^{(1)}_{0l}(c,\xi) \qquad (7)$$

$$he_{0l}(c,\xi) = R^{(1)}_{0l}(c,\xi) + iR^{(2)}_{0l}(c,\xi) \qquad (8)$$

and $\mathrm{d}^2x_i = (\xi_i^2 - \eta_i^2)\, \mathrm{d}\xi_i\, \mathrm{d}\eta_i$. Further defining the two auxiliary integrals,

$$K^{(1)}_{pqrsl}(k) = \int \mathrm{d}^2x_1\, \mathrm{d}^2x_2\, S^{(1)}_{0l}\left(\frac{k}{2},\eta_1\right) S^{(1)}_{0l}\left(\frac{k}{2},\eta_2\right) R^{(1)}_{0l}\left(\frac{k}{2},\xi_<\right) R^{(1)}_{0l}\left(\frac{k}{2},\xi_>\right) f_{pqrs}(\xi_1,\eta_1,\xi_2,\eta_2) \qquad (9)$$

and

$$K^{(2)}_{pqrsl}(k) = \int \mathrm{d}^2x_1\, \mathrm{d}^2x_2\, S^{(1)}_{0l}\left(\frac{k}{2},\eta_1\right) S^{(1)}_{0l}\left(\frac{k}{2},\eta_2\right) R^{(1)}_{0l}\left(\frac{k}{2},\xi_<\right) R^{(2)}_{0l}\left(\frac{k}{2},\xi_>\right) f_{pqrs}(\xi_1,\eta_1,\xi_2,\eta_2) \qquad (10)$$

Equation 5 takes the form

$$g_{pqrs}(k) = 4\pi^2 ik \sum_{l=0}^{\infty} (2l+1)[K^{(1)}_{pqrsl}(k) + i\, K^{(2)}_{pqrsl}(k)] \qquad (11)$$

To apply eq 3, we expand eq 11 in powers of k using the power expansions of the spheroidal wave functions. A useful expansion for the angular functions of the first kind is[18]

$$S^{(1)}_{0l}(k/2, z) = d^{0l}_0(k) \sum_{j=0}^{\infty}\left[\sum_{k=b_{-2j,k_0}}^{2j}{}' \alpha^{0l}_{jk} P_{l+k}(z)\right]\left(\frac{k}{2}\right)^{2j} \qquad (12)$$

where $\sum'$ means that the sum is over even values of the index, $P_{l+k}(z)$ are the Legendre functions,[20] $b_{ij} = \max(i,j)$, $k_0 = l \bmod(2) - l$, and the recursive relations for the coefficients α^{0l}_{jk} are given in the appendix. In ref 18, it is also shown that $d^{0l}_0(k)$ admits the expansion

$$[d^{0l}_0(k/2)]^2 = \left[\sum_{j1=0}^{\infty}\sum_{j2=0}^{\infty}\sum_{k=-a_{2j1,2j2}}^{\alpha_{2j1,2j2}}{}' \alpha^{0l}_{j1k}\,\alpha^{0l}_{j2k}\frac{2l+1}{2l+2k+1}\left(\frac{k}{2}\right)^{2(j1+j2)}\right]^{-1} \qquad (13)$$

where $a_{i,j} = \min(i,j)$. The other special functions admit similar expansions, and we do not reproduce the details here. Then, we obtain for the auxiliary integrals

$$K^{(1)}_{pqrsl}(k) = \frac{(l!)^2}{4^{2l+1}}\left[\frac{\sum_{j=0}^{\infty}\delta_{jl}(k/2)^{2j}}{\sum_{j,i=0}^{\infty}\Delta_{ijl}(k/2)^{2(i+j)}}\right]^2\left[\frac{1}{\sum_{j,i=0}^{\infty}\tau^l_{ij}(k/2)^{2(i+j)}}\right] \sum_{j1\,j2,j3,j4=0}^{\infty}(k/2)^{2(j1+j2+j3+j4+l)} \int \mathrm{d}^2x_1\, \mathrm{d}^2x_2\, \gamma^{(1)}_{lj1}(\eta_1)\, \gamma^{(1)}_{lj2}(\eta_2)\, \gamma^{(1)}_{lj3}(\xi_<)\, \gamma^{(1)}_{lj4}(\xi_>)\, f_{pqrs}(\xi_1,\eta_1,\xi_2,\eta_2) \qquad (14)$$

where

$$\tau^l_{ij} = \sum_{k=-a_{2i,2j}}^{\alpha_{2i,2j}}{}' \alpha^{0l}_{ik}\,\alpha^{0l}_{jk}\frac{2l+1}{2l+2k+1} \qquad (15)$$

$$\delta_{jl} = \sum_{k=0}^{j} \frac{\alpha^{0l}_{j,-2k}}{k!\Gamma(l-k+3/2)} \qquad (16)$$

$$\Delta_{ijl} = \sum_{k=0}^{i}\sum_{m=-2j}^{2j}{}' \frac{\alpha^{0l}_{i,2k}\,\alpha^{0l}_{jm}}{k!\Gamma(-l-k+1/2)} \qquad (17)$$

and

$$\gamma_{lj}^{(1)}(z) = \sum_{k=b_{-2j,k_0}}^{2j\,\prime} \alpha_{jk}^{0l}\, P_{l+k}(z) \tag{18}$$

Analogously, $K_{pqrsl}^{(2)}(k)$ may also be expanded as

$$K_{pqrsl}^{(2)} = -\frac{2}{k}\left[\frac{\sum_{j,i=0}^{\infty} \delta_{j,l}\tilde{\delta}_{i,l}(k/2)^{2(i+j)}}{\sum_{i1,i2,i3,i4=0}^{\infty} \Delta_{i2,i1,l}\tilde{\Delta}_{i4,i3,l}(k/2)^{2(i1+i2+i3+i4)}}\right]$$
$$\left[\frac{1}{\sum_{j,i=0}^{\infty} \tau_{ij}^{l}(k/2)^{2(i+j)}}\right] \sum_{j1,j2,j3,j4=0}^{\infty} (k/2)^{2(j1+j2+j3+j4)} \int \mathrm{d}^2x_1\, \mathrm{d}^2x_2$$
$$\gamma_{lj1}^{(1)}(\eta_1)\, \gamma_{lj2}^{(1)}(\eta_2)\, \gamma_{lj3}^{(1)}(\xi_<)\, \gamma_{lj4}^{(2)}(\xi_>)\, f_{pqrs}(\xi_1,\eta_1,\xi_2,\eta_2) \tag{19}$$

where

$$\tilde{\delta}_{jl} = \sum_{k=0}^{j} \frac{\alpha_{j,2k}^{0l}}{k!\Gamma(-l-k+1/2)} \tag{20}$$

$$\tilde{\Delta}_{ijl} = \sum_{k=0}^{i} \sum_{m=-2j}^{2j\,\prime} \frac{\alpha_{i,-2k}^{0l}\, \alpha_{jm}^{0l}}{k!\Gamma(l-k+3/2)} \tag{21}$$

and

$$\gamma_{lj}^{(2)}(z) = \sum_{k=-2j}^{a_{k_0-2,-2j}\,\prime} \tilde{\alpha}_{jk}^{0l}\, P_{-l-k-1}(z) + \sum_{k=b_{k_0,-2j}}^{2j\,\prime} \alpha_{jk}^{0l}\, Q_{k+l}(z) \tag{22}$$

where $Q_{k+l}(z)$ are the Legendre functions[20] and the recursive expressions for the coefficients $\tilde{\alpha}_{jk}^{0l}$ are shown in the appendix.

Note that integrals present in eqs 14 and 19 do not depend on k. There are three different integrals

$$B_{m,q} = \int_{-1}^{1} P_m(\eta)\, \eta^q\, \mathrm{d}\eta \tag{23}$$

$$Z_{mnlk}^{\xi}(\alpha,\beta) = \int \mathrm{d}\xi_1\, \mathrm{d}\xi_2\, P_m(\xi_<)\, P_n(\xi_>)\, \xi_1^l\, \xi_2^k\, \mathrm{e}^{-\alpha\xi_1}\, \mathrm{e}^{-\beta\xi_2} \tag{24}$$

and

$$W_{mnlk}^{\xi}(\alpha,\beta) = \int \mathrm{d}\xi_1\, \mathrm{d}\xi_2\, P_m(\xi_<)\, Q_n(\xi_>)\, \xi_1^l\, \xi_2^k\, \mathrm{e}^{-\alpha\xi_1}\, \mathrm{e}^{-\beta\xi_2} \tag{25}$$

Integrals 23 and 25 were solved by McEachran and Cohen.[3] They obtained analytic recursion formulas sufficient to generate these integrals. For solving the integrals in eq 24, which are similar to the integrals in eq 25, we used the scheme presented in ref 3.

Now, we arrive at a final expression for $g_{pqrs}(k)$:

$$g_{pqrs}(k) = 8\pi^2 \sum_{lkentN=0}^{\infty} (2l+1)\left[\frac{i(l!)^2}{4^{2l+1}} h_{2N}^{l}\, J_{kentpqrs}^{l}\left(\frac{k}{2}\right)^{2(k+e+n+t+l+N)+1} + \tilde{h}_{2N}^{l}\, Y_{kentpqrs}^{l}\left(\frac{k}{2}\right)^{2(k+e+n+t+N)}\right] \tag{26}$$

where

$$J_{kentpqrs}^{l} = \sum_{k1=b_{-2k,k_0}}^{2k\,\prime} \sum_{k2=b_{-2e,k_0}}^{2e\,\prime} \sum_{k3=b_{-2n,k_0}}^{2n\,\prime} \sum_{k4=b_{-2t,k_0}}^{2t\,\prime} A_{k1,k2,k3,k4}(l,k,e,n,t)\, B_{l+k1,q}[Z_{l+k3,l+k4,p,r+2}^{\xi}(\alpha,\beta)\, B_{l+k2,s} - Z_{l+k3,l+k4,p,r}^{\xi}(\alpha,\beta)\, B_{l+k2,s+2}] \tag{27}$$

and

$$Y_{kentpqrs}^{l} = \sum_{k1=b_{-2k,k_0}}^{2k\,\prime} \sum_{k2=b_{-2e,k_0}}^{2e\,\prime} \sum_{k3=b_{-2n,k_0}}^{2n\,\prime} \sum_{k4=-2t}^{a_{k_0-2,2t}\,\prime} \tilde{A}_{k1,k2,k3,k4}(l,k,e,n,t)\, B_{l+k1,q}[Z_{l+k3,-l-k4-1,p,r+2}^{\xi}(\alpha,\beta)\, B_{l+k2,s} - Z_{l+k3,-l-k4-1,p,r}^{\xi}(\alpha,\beta)\, B_{l+k2,s+2}] + \sum_{k1=b_{-2k,k_0}}^{2k\,\prime} \sum_{k2=b_{-2e,k_0}}^{2e\,\prime} \sum_{k3=b_{-2n,k_0}}^{2n\,\prime} \sum_{k4=b_{-2t,k_0}}^{2t\,\prime} A_{k1,k2,k3,k4}(l,k,e,n,t)\, B_{l+k1,q}[W_{l+k3,l+k4,p,r+2}^{\xi}(\alpha,\beta)\, B_{l+k2,s} - W_{l+k3,l+k4,p,r}^{\xi}(\alpha,\beta)\, B_{l+k2,s+2}] \tag{28}$$

where the coefficients A are

$$A_{k1,k2,k3,k4}(l,k,e,n,t) = \alpha_{k,k1}^{0l}\, \alpha_{e,k2}^{0l}\, \alpha_{n,k3}^{0l}\, \alpha_{t,k4}^{0l}$$
$$\tilde{A}_{k1,k2,k3,k4}(l,k,e,n,t) = \alpha_{k,k1}^{0l}\, \alpha_{e,k2}^{0l}\, \alpha_{n,k3}^{0l}\, \tilde{\alpha}_{t,k4}^{0l} \tag{29}$$

The coefficients h_N^l and $\tilde{h}_N^l$ in eq 26 are obtained from algebraic treatment of eqs 14 and 19 by just grouping terms in powers of k.

We can see that, when we introduce eq 26 in eq 3, all the series become finite. When m in eq 1 is odd, $m+1$ in eq 3 is even and just the second term in eq 26 survives. In this case, the sum over k, e, n, t, and N is truncated by the condition $2(k+e+n+t+N) = m+1$. For the sum over l, we have to analyze eq 28. It is straightforward to show that $B_{mq} = 0$ for $m > q$; then, the sum over l is truncated by the condition $l + k_1 \leq q$. For even values of m, $m+1$ is odd, just the first term in eq 26 survives, and the sum over k, e, n, t, N, and l is truncated by the condition $2(k+e+n+t+N+l) = m$.

3. Numerical Discussion

The iterative method presented in section 2 and its application to the variational calculation of the ground state of the hydrogen molecule have been tested in extensive numerical computations. It is interesting to discuss some numerical problems. The first one is the use of analytical recursion relations. It is known that these relations are numerically very unstable, to the extreme that one or two significant

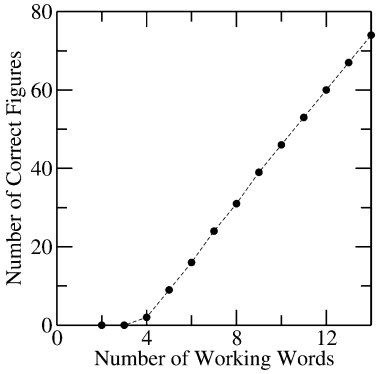


Figure 1. Number of correct figures against number of working words for the integral I^m_{pqrs} with $\alpha = \beta = 1.2$, $p = 12$, $q = 12$, $r = 10$, $s = 10$, and $m = 7$.

figures may be lost by iteration. To avoid this problem, we wrote all our codes using MPFUN, a multiprecision Fortran-90 package,[21] which allows for working with an arbitrary precision. MPFUN was successfully applied to high-precision calculations in quantum few-body systems.[22,23] In this work, we made our calculations with 100 figures, in contrast to the maximum 32 figures that allows quadruple precision in standard Fortran-90.

To check the numerical stability of our method, we evaluated one of the typical integrals in eq 1, I^m_{pqrs}, with different accuracy levels. In particular, in Figure 1, we show the number of correct figures obtained for the integral I^m_{pqrs} with $p = 12$, $q = 12$, $r = 10$, $s = 10$, and $m = 7$ as a function of the used accuracy level in words (one word is equal to seven figures).[21] The graphic is almost independent of the value of m, and the numerical error grows with p, q, r, and s; therefore, we chose their maximum values used in our calculations in section 4. We varied the accuracy from 1 to 14 words ($\simeq$100 digits), and we compared it with the result obtained using 42 ($\simeq$300 digits), which is considered to have more than 100 correct figures. It is interesting to note that no correct digits are obtained with three-word calculations. This means that it is not possible to use standard double precision Fortran (16 figures). We get only nine correct digits with five words ($\simeq$35 figures); therefore, no reliable results may be obtained using standard quadruple precision (32 figures). In Table 1, integrals with 70 significant figures are presented.

Once the integrals are calculated, they are used in Ritz variational calculations (see section 4). The James−Coolidge basis set is not orthogonal; for this reason, it is necessary to solve a generalized eigenvalue problem.[24] The solution of a linear system with ill-conditioned matrices produces a significant loss of numerical accuracy which has to be added to the accuracy lost in the first step of the work (integrals evaluation). In our case, all the matrices involved in the generalized eigenvalue problem are extremely ill-conditioned. The overlap matrix is a positive definite matrix, but it may become nonpositive as a result of numerical accuracy problems. To avoid this, it is necessary to compute the matrix elements with great accuracy. Frolov and Bailey,[22] in the study of three body systems, ensure that they needed to work with 84−100 digits in order to produce final results with 30 correct figures. In this work, we started the calculations with 100 figures to obtain all the matrix elements with 70 correct figures. The generalized eigenvalue problem is solved applying the *Cholesky decomposition* to the overlap matrix in order to recover a standard symmetric eigenvalue problem.[24] This transformation was also done with the MPFUN package. Although the CPU time for high-precision calculation is expected to scale as a power of the number of working words, the integrals were evaluated in a reasonable time on a personal computer. The numerical evaluation of all the integrals used in section 4 takes about 81 min and the Cholesky decomposition 155 min on a 3 GHz Pentium 4 processor.

The transformed matrix elements have more than 16 correct figures. Then, the last step, the eigenvalues determination, is performed in standard double-precision Fortran-90.

4. The Hydrogen Molecule

As an application of the method described in section 2, we calculate the ground-state energy and equilibrium radii of the hydrogen molecule in the Born−Oppenheimer approximation. The Hamiltonian of this system, in atomic units, is

$$H = \sum_{i=1}^{2}\left(-\frac{1}{2}\nabla_i^2 - \frac{1}{|\vec{r}_i - \vec{R}/2|} - \frac{1}{|\vec{r}_i + \vec{R}/2|}\right) + \frac{1}{r_{12}} + \frac{1}{R} \quad (30)$$

To apply the Ritz variational principle, we need to evaluate where

$$H_{ij} = \langle\phi_i| H|\phi_j\rangle \qquad S_{ij} = \langle\phi_i|\phi_j\rangle \quad (31)$$

$$\phi_n = C\, e^{-\alpha(\xi_1+\xi_2)}\,(\xi_1^{p_n}\,\eta_1^{q_n}\,\xi_2^{r_n}\,\eta_2^{s_n} + \xi_1^{r_n}\,\eta_1^{s_n}\,\xi_2^{p_n}\,\eta_2^{q_n}) r_{12}^{m_n} \quad (32)$$

Here, C is the normalization constant. It is obvious that the ground-state wave function has to be invariant under inversion with respect to the plane of symmetry of the molecule. As a result, we have the restriction that $q_n + s_n$ must be even. To obtain the ground-state energy of the system, we have to find the lowest root of

$$\mathrm{Det}(\mathbf{H} - E\mathbf{S}) = 0 \quad (33)$$

An optimization of the parameter α was done with a 1710-term wave function. The optimal value obtained was $\alpha = 1.2$.

In this work, the quantum numbers of the basis function eq 32 are allowed to take values from 0 to 5. We calculated the ground-state energy of the hydrogen molecule for different values of the internuclear distance R with the 2052-term wave function.

We obtained for the equilibrium distance $R_{eq} = 1.401\,08$ and for the correspondent ground-state energy $E_0(R_{eq}) = -1.174\,475\,930\,2$ au. In Table 2, we show the energy for $R = 1.4$ as we increase the correlation power m. The case $m = 0$ is the noncorrelated approximation.[3] The values for $R = 1.4$ were calculated for comparison with other results available in the literature. Our value for the ground-state energy of the hydrogen molecule for $R = 1.4$ is lower than the values reported by Kolos, using a modified James−Coolidge expansion with two variational parameters,[4] and

Table 1. Integrals I^m_{pqrs} for $\alpha = \beta = 1.2$ and $p = 12$, $q = 12$, $r = 10$, $s = 10$, $m = -1$ to 10

m	$I^m_{12,12,10,10}$ ($\alpha = \beta = 1.2$)
−1	$1.57300119313815737534218727783205609513409646164423018824196943556488 \times 10^7$
0	$3.17633036827265043005880147966246050951855666030622938765912707865055 \times 10^7$
1	$1.09074258787096404795583398355214154417898838939751969077654689628259 1 \times 10^8$
2	$5.02172470737598712528444255110052932666774039975792403476588900650831 0 \times 10^8$
3	$2.65817477568976661283202969841275889311841204355893144618125446538961 3 \times 10^9$
4	$1.51542676526141929983359175647088621523206491315769271456971041157278 1 \times 10^{10}$
5	$9.07751712865339763972791420535719399176438997006700743254874665374165 3 \times 10^{10}$
6	$5.65735664424462941895866058785091086833457956462022985390578850848976 1 \times 10^{11}$
7	$3.65209422280077022706293063568996178374516785863862566056671385555171 6 \times 10^{12}$
8	$2.43616708669377733636455422545478717182789437468049229455776847334140 9 \times 10^{13}$
9	$1.67659769602840474744254782186841761090515875645837010613785395536628 3 \times 10^{14}$
10	$1.18903123442032702375131635468010562531972265946810869902441000487831 1 \times 10^{15}$

Table 2. Variational Ground-State Energy for the Hydrogen Molecule with Maximum Values $p = 5$, $q = 5$, $r = 5$, and $s = 5$

m	$E(R = 1.4)$
0	−1.161 482 984 33
1	−1.174 435 130 87
2	−1.174 475 155 80
3	−1.174 475 700 00
4	−1.174 475 711 87
5	−1.174 475 713 00
W. Kolos (1994)[4]	−1.174 475 686
H. Nakatsuji (2004)[5]	−1.174 475 703

by Nakatsuji, using a modified James−Coolidge expansion allowing negatives values of (p_n, r_n) in the basis function eq 32.

5. Conclusions

The main contribution of this paper is a new analytical method for the evaluation of the James−Coolidge two-center, two-electron integrals. The method is based on the standard expansion for the free Green function for the Helmholtz operator in spheroidal wave functions[17] and uses the new expressions obtained by Falloon for those special functions.[18,19]

The formulas presented have been successfully tested in numerical calculations for the Born−Oppenheimer hydrogen molecule ground-state energy. To obtain the Hamiltonian matrix elements correct up to 16 decimal places, we used the MPFUN package[21] with roughly 100 digits in the recurrence relations for the integrals and Cholesky decomposition of the overlap matrix.

The method presented in this work is appropriate for high-precision variational calculations of bound states of other two-center, two-electron Hamiltonians. As a relevant application, the existence of dipole dianions will be addressed in a forthcoming paper.

Acknowledgment. We thank A. Banchio for a critical reading of the manuscript; this work has been supported by CONICET, SECYT-UNC, and Agencia Córdoba Ciencia.

Appendix

Recurrence Relations for the Coefficients. Here, we present the recurrence relation for the coefficients α^{0l}_{jk} and $\tilde{\alpha}^{0l}_{jk}$ taken from ref 18, which are

$$\alpha^{0l}_{jk} = \frac{1}{(l+k)(l+k+1) - l_0^{0l}}\Big[\sum_{i=0}^{j-1} l^{0l}_{i+1}\,\alpha^{0l}_{j-i-1,k} - (a_{0lk}\alpha^{0l}_{j-1,k+2} + \beta_{0lk}\alpha^{0l}_{j-1,k} + \gamma_{0lk}\alpha^{0l}_{j-1,k-2})\Big] \quad (34)$$

$$a_{0lk} = \frac{(l+k+1)(l+k+2)}{(2l+2k+3)(2l+2k+5)} \quad (35)$$

$$\beta_{0lk} = \frac{1}{2}\left(1 + \frac{1}{(2l+2k-1)(2l+2k+3)}\right) \quad (36)$$

$$\gamma_{0lk} = \frac{(l+k)(l+k-1)}{(2l+2k-3)(2l+2k-1)} \quad (37)$$

$$l_0^{0l} = l(l+1), \qquad l_1^{0l} = \beta_{0l0}, \qquad l_j^{0l} = a_{0l0}\alpha^{0l}_{j-1,2} + \gamma_{0l0}\alpha^{0l}_{j-1,-2} \quad j \geq 2 \quad (38)$$

$$\tilde{\alpha}^{0l}_{jk} = \frac{1}{(l+k)(l+k+1) - l_0^{0l}}\Big[\sum_{i=0}^{j-1} l^{0l}_{i+1}\,\tilde{\alpha}^{0l}_{j-i-1,k} - (\tilde{a}_{0l}\alpha^{0l}_{j-1,k+2} + \beta_{0lk}\tilde{\alpha}^{0l}_{j-1,k} + \gamma_{0lk}\tilde{\alpha}^{0l}_{j-1,k-2})\Big] \quad (39)$$

for $k = k_0 - 2$ and

$$\tilde{\alpha}^{0l}_{jk} = \frac{1}{(l+k)(l+k+1) - l_0^{0l}}\Big[\sum_{i=0}^{j-1} l^{0l}_{i+1}\,\tilde{\alpha}^{0l}_{j-i-1,k} - (a_{0l}\alpha^{0l}_{j-1,k+2} + \beta_{0lk}\tilde{\alpha}^{0l}_{j-1,k} + \gamma_{0lk}\tilde{\alpha}^{0l}_{j-1,k-2})\Big] \quad (40)$$

for $k = k_0 - 4$, $k_0 - 6$, and so forth, and $\tilde{a}_{0l} = 1$ for even l values, and $\tilde{a}_{0l} = 1/3$ for odd l values.

References

(1) James, H. M.; Coolidge, A. S. *J. Chem. Phys.* **1933**, *1*, 825−835.

(2) Kolos, W.; Roothaan, C. C. J. *Rev. Mod. Phys.* **1960**, *32*, 219−229.

(3) McEachran, R. P.; Cohen, M. *Isr. J. Chem.* **1975**, *13*, 5−13.

(4) Kolos, W. *J. Chem. Phys.* **1994**, *101*, 1330−1332.

(5) Nakatsuji, H. *Phys. Rev. Lett.* **2004**, *93*, 030403.

(6) Skurski, P.; Gutowski, M.; Simons, J. *Int. J. Quantum Chem.* **2000**, *76*, 197−204.

(7) Sarasola, C.; Fowler, J.; Elorza, J. M.; Ugalde, J. M. *Chem. Phys. Lett.* **2001**, *337*, 355−360.

(8) Dreuw, A.; Cederbaum, L. S. *Chem. Rev.* **2002**, *102*, 181−200.

(9) Trindle C.; Yumak, A. *J. Chem. Theory Comput.* **2005**, *1*, 433−438.

(10) Jensen, A.S.; Riisager, K.; Fedorov, D. V.; Garrido, E. *Rev. Mod. Phys.* **2004**, *76*, 215−261.

(11) Serra, P.; Kais, S. Manuscript in preparation.

(12) Kais, S.; Serra, P. *Adv. Chem. Phys.* **2003**, *125*, 1−99.

(13) Serra, P.; Kais, S. *Chem. Phys. Lett.* **2003**, *372*, 205−209.

(14) Ferrón, A.; Serra, P.; Kais, S. *J. Chem. Phys.* **2004**, *120*, 8412−8419.

(15) Hylleraas, E. A. *Z. Phys.* **1929**, *54*, 347.

(16) Kolos, W.; Roothaan, C. C. J. *Rev. Mod. Phys.* **1960**, *32*, 205−210.

(17) Morse, P. M.; Feshbach, H. *Methods of Theoretical Physics*; McGraw-Hill: New York, 1953; Vol. 1 and Vol. 2.

(18) Falloon, P. E. Theory and Computation of Spheroidal Harmonics with General Arguments. Ph.D. Thesis, University of Western Australia, Crawley, Western Australia, September 2001.

(19) Falloon, P. E.; Abbott, P. C.; Wang, J. B. *J. Phys. A.* **2003**, *36*, 5477−5495.

(20) Abramowitz, M.; Stegun, I. A. *Handbook of Mathematical Functions*; Dover: New York, 1972.

(21) Bailey, D. H. *ACM Trans. Math. Soft.* **1995**, *21*, 379−387.

(22) Bailey, D. H.; Frolov, A. M. *J. Phys. B* **2002**, *35*, 4287−4297.

(23) Frolov, A. M.; Bailey, D. H. *J. Phys. B* **2003**, *36*, 1857−1867.

(24) Press, W.; Teukolsky, S.; Vetterling, W.; Flannery, B. *Numerical Recipes in Fortran 77*, 2nd edition; Cambridge University Press: Cambridge, 1992.

CT0502662

JCTC Journal of Chemical Theory and Computation

The ω, ϕ, and ψ Space of *N*-Hydroxy-*N*-methylacetamide and *N*-Acetyl-*N*′-hydroxy-*N*′-methylamide of Alanine and Their Boron Isosteres

Alpeshkumar K. Malde, Santosh A. Khedkar, and Evans C. Coutinho*

Department of Pharmaceutical Chemistry, Bombay College of Pharmacy, Kalina, Santacruz (E), Mumbai 400 098, India

Received September 28, 2005

Abstract: The conformational space of *N*-hydroxy-*N*-methylacetamide [$CH_3-CO-N(OH)CH_3$, NMAOH] and its boron isostere [$CH_3-CO-B(OH)CH_3$, BMAOH] has been studied by quantum chemical methods. The potential energy surface of NMAOH and BMAOH has been built at the HF, B3LYP, and MP2 levels of theory with the 6-31+G* basis set. The minima and transition states for rotations about various torsional angles have been located, and the energy barriers have been estimated. The global minimum energy structure of both peptides exhibits an intramolecular hydrogen bond between the carbonyl oxygen and the hydroxyl group, imparting a conformational rigidity to the peptides. The *omega* rotation barrier is lower in the boron isostere than in NMAOH. The difference in the rotation barrier has been attributed to second-order orbital interactions, like negative hyperconjugation, as revealed by NBO calculations. In contrast, the rotation barrier around the torsion angle *tau* (torsion governing rotation about the N−O and B−O bonds) is relatively higher in the boron analogue. This difference is due to the double bond character in the B−O bond as opposed to the N−O bond which has the character of a single bond. As an extension, *N*-acetyl-*N*′-hydroxy-*N*′-methylamide of alanine (Ala-NOH) and its boron isostere (Ala-BOH) have been adopted as model peptides to study the conformational preferences about the ϕ and ψ torsion angles. The study reveals a strong preference for a Type I beta turn as well as inclinations for a left-handed alpha helix, for positive *phi* torsions, and for extended *psi* conformations for Ala-NOH; Ala-BOH, on the other hand, shows a leaning toward positive *phi* and extended *psi*, with no preference for any regular secondary structure motifs. The replacement of nitrogen by boron changes the electronic and conformational properties of the peptide, extending greater flexibility around the *omega* angle, a strong preference for positive *phi* values, and a shift in the site of nucleophilic attack from the carbonyl group to boron.

Introduction

Peptides and proteins are one of the important classes of biomolecules. The conformations of peptides and protein are crucial determinants of their biological effects. The values of the three backbone torsion angles—*omega* (ω), *phi* (ϕ), and *psi* (ψ)—dictate the secondary structure of peptides.[1] Most natural peptides adopt ω with 180° (trans), and occasionally, ω assumes 0° (cis) for peptides with the Xxx-Pro and Xxx-Gly motifs.[2] The ϕ and ψ values in natural peptides and proteins are restricted to the allowed regions of the Ramachandran space.[1] Peptides form an important area of therapeutics[3], e.g. insulin, substance P, growth hormone, thyrotropin releasing hormone, gastric inhibitory polypeptide, gastrin, neurokinins, bradykinin, etc. have important therapeutic applications. There are certain advantages with peptide

* Corresponding author phone: +91-22-26670871; fax: +91-22-26670816; e-mail: evans@bcpindia.org.

10.1021/ct050242v CCC: $33.50
Published on Web 01/05/2006

therapeutics. Hormones and neurotransmitter peptides are very potent and consequently administered in very small doses, besides exhibiting a high selectivity and specificity in binding to their target. However, there are also complexities in using and developing peptides as therapeutics. The oral delivery of peptides is restricted due to degradation at the "scissile" amide bond. Imparting potency, specificity, and selectivity for peptides designed from natural analogues for certain biological end-points still remains a challenge. Issues such as proteolytic stability, potency, specificity, and selectivity can be addressed by modification of the amide bond and/or isosteric/bioisosteric replacement of the backbone atoms of the peptide. *N*-methylation;[4] *N*-hydroxylation;[5] replacement of the amide bond[6,7] by sulfonamide, phosphonamidate, and carbamate; isosteric replacement of the carbonyl carbon with boron (peptide boronic acid[8,9]); and isosteric replacement of the alpha carbon with boron (ammonia-carboxyboranes[10−12]) have been reported in the literature as techniques to explore new peptide conformations and to design "druglike," proteolytically stable molecules. *N*-Hydroxylation of peptides has been used to impart conformational rigidity through formation of an intramolecular hydrogen bond with the CO group.[5] This imparts an ability to chelate metal ions for specific binding to proteins containing metals in the active site.

We had reported for the first time a boron isostere of the amide nitrogen in peptides and studied the ω, ϕ, and ψ preferences by ab initio and density functional methods.[13,14] These molecules were designed as plausible serine protease inhibitors. The replacement of nitrogen with boron leads to two new characteristics: a preference for the ω angle for 90°, in contrast to 180° or 0° for natural peptides, and second, conformations that lie in the "disallowed regions" (positive ϕ angles) of the Ramachandran plot. These peptides also exhibit greater flexibility around the ω angle.

In this paper, we look at the ω, ϕ, and ψ preferences of an *N*-hydroxy peptide and its boron isostere, by ab initio and density functional methods. *N*-Methylacetamide (NMA, **I**) has been extensively studied, both experimentally and theoretically, as a model for the peptide backbone. In a like manner, *N*-methyl-*N*-hydroxyacetamide (NMAOH, **II**) is a good model to study *N*-hydroxy peptides and acetylmethylhydroxyborane (BMAOH, **III**) an analogous model for the boron isostere. In addition, *N*-acetyl-*N*′-hydroxy-*N*′-methylamide of alanine (Ala-NOH, **IV**) and its boron isostere (Ala-BOH, **V**) have been adopted as models to study the ϕ and ψ distribution of such peptides. The hypersurfaces of NMAOH (**II**) and BMAOH (**III**), with its associated ground and transition states, and the corresponding ground states of Ala-NOH (**IV**) and Ala-BOH (**V**) have been mapped by ab initio Hartree−Fock (HF), density functional, and post-HF methods. Second-order orbital interactions by Natural Bond Orbitals (NBO) method were also carried out to understand the fundamental differences in the structures of the *N*-hydroxy peptides and their boron isosteres.

Computational Details

Ab initio molecular orbital[15] and density functional theory[16] calculations have been carried out using the Gaussian03W[17] (revision C.01) package running on a Pentium III processor with 512 MB RAM. The stability of all wave functions was checked at the HF,[18] Becke's three parameter exchange functional, and the gradient corrected functional of Lee, Yang, and Paar (B3LYP),[19−21] second-order Møller−Plesset MP2 (full)[22,23] level of theory using the 6-31+G* basis set.

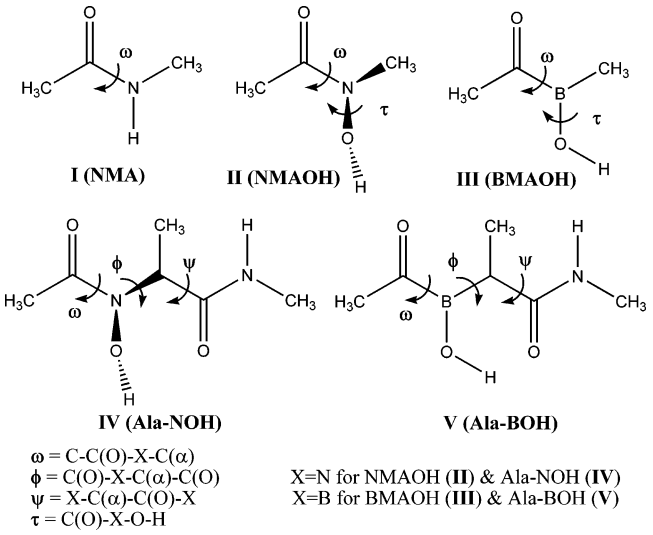
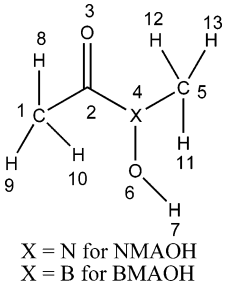


Figure 1. Structures of NMA, NMAOH, and Ala-NOH and their boron counterparts.

Chart 1



The atom labels for NMAOH (**II**) and BMAOH (**III**) are listed in Chart 1, and the two torsion angles, ω and τ, are defined as shown in Figure 1. In NMAOH, the hydroxylamine moiety can adopt two conformations around the N−O bond. In the first, the two lone pairs of electrons of O are *syn-clinal* and in the second, *anti-clinal* with respect to the lone pairs of electrons on N. This has been observed from a conformational search of hydroxylamine by ab initio calculations. These initial two conformations around the N−O bond in NMAOH were chosen, and for each arrangement of τ, a scan in increments of 30° of the ω torsion angle was carried out at the HF/6-31+G* level of theory. Conformations with an ω value of 30° and 210° were found to be the lowest in energy. Now, for each of these two conformations with ω values of 30° and 210°, respectively, a τ scan in increments of 30° was run at the HF/6-31+G* level of theory. The minima saddle points for rotation about the ω torsion and saddle points for rotation about the τ torsion rotation were thus identified. All these conformations were further optimized at the B3LYP and MP2 levels of theory with the same basis set, and the conformations were confirmed by frequency calculations, which returned one imaginary frequency for

Table 1. Relative Energies (kcal/mol) of Various Minima and Transition States on PES of NMAOH (**II**) at the HF, B3LYP, and MP2 Levels of Theory with the 6-31+G* Basis Set[c]

				HF/6-31+G*	B3LYP/6-31+G*	MP2(full)/6-31+G*		
		NIMAG	PG	rel.[a]	rel.[a]	rel.[a]	ω^b	τ^b
minima	GM	0	C_1	0.0	0.0	0.0	32	10
	LM	0	C_1	1.8	1.1	0.4	202	120
ω rotation transition state (TS)	ωTS1/ωTS1′	1	C_1	13.6	15.6	12.6	125/−125	123/−123
	ωTS2/ωTS2′	1	C_1	13.5	15.3	13.0	135/−135	−60/60
	ωTS3/ωTS3′	1	C_1	14.1	15.8	13.3	36/−36	−105/105
	ωTS4/ωTS4′	1	C_1	21.1	21.7	20.3	−39/39	−74/74
τ rotation TS	τTS1	1	C_1	13.2	12.2	12.2	39	−146
	τTS2	1	C_1	8.5	6.5	7.3	−165	7
	τTS3	1	C_1	6.5	6.2	6.3	−159	−147
pyramidal inversion TS	PyTS	1	C_s	2.0	0.0	0.4	0	0

[a] Relative energy in kcal/mol. [b] Torsion angle in degrees. [c] NIMAG = number of imaginary frequency, PG = point group, GM = global minimum, LM = local minimum.

each transition state and all positive frequencies for each ground state.

A similar strategy was adopted for probing the conformational space of BMAOH. The hydroxylborane moiety has a planar conformation, and the resulting τ angles in BMAOH are either 0° or 180°. The two BMAOH conformations with τ values of 0° and 180° were then examined by an ω scan in increments of 30° at the HF/6-31+G* level of theory. Two conformations with ω values of 0° and 180° were identified as the lowest in energy. These two conformations with ω values of 0° and 180° were then evaluated by a τ scan in increments of 30° at the HF/6-31+G* level of theory. The minima, saddle points for rotation around the ω angle, and saddle points for rotation about the τ angle were thus located. All these structures were further optimized at the B3LYP and MP2 levels of theory using the 6-31+G* basis set and confirmed by frequency calculations.

The NBO[24−26] analysis was carried out on the minimum energy structures of NMAOH (**II**) and BMAOH (**III**), optimized at the MP2(full)/6-31+G* level, to quantitatively estimate the second-order interactions as $E_{ij} = -2F_{ij}/\Delta E_{ij}$, where E_{ij} is the energy of the second-order interaction; $\Delta E_{ij} = E_i - E_j$ is the energy difference between the interacting molecular orbitals i and j; and F_{ij} is the Fock matrix element for the interaction between orbitals i and j. The "atomic partial charges" of the global minimum of NMAOH (**II**) and BMAOH (**III**), optimized at the MP2(full)/6-31+G* level, were calculated using Natural Population Analysis (NPA) as implemented in NBO and additionally by the 'ESP fit' method formulated by Merz, Singh, and Kollman.[27]

For Ala-NOH (**IV**), the minima in the ω and τ space was searched starting with two different conformations for ω and τ as identified previously for NMAOH (**II**). This corresponds to structures with $\omega = 32°$; $\tau = 10°$ and $\omega = 202°$; $\tau = 120°$. For each (ω, τ) pair, 144 conformations were generated with 30° increments of the ϕ, ψ dihedrals. Each conformation was geometry optimized first at the HF/3-21G level of theory with "constraints" on the initial ϕ, ψ angles. A Ramachandran plot of the 144 conformations was constructed, and conformations within 5.0 kcal/mol of the global minimum were identified. These low-energy conformations were further optimized without constraints at the B3LYP/6-31+G* level of theory. A similar study was carried out for Ala-BOH (**V**) with the starting (ω, τ) pairs of (0°, 0°) and (153°, 180°).

Results and Discussion

All wave functions for molecules **II**−**V** were found to be stable under the perturbations considered at the HF, B3LYP, and MP2 levels of theory.

Potential Energy Surface (PES) of NMAOH (II). For NMAOH (**II**) besides the global minimum (GM), there is also a local minimum (LM) within 2.0 kcal/mol of the GM. For each structure, several transition states (TS) for rotation about the ω angle exists. The geometries of these TS depend on the state of the pyramidal amide nitrogen, i.e. the lone pair of electrons on nitrogen may either be directed downward, which we label as 'pyramidal up', or the lone pair of electrons on nitrogen may be positioned upward, which we call as 'pyramidal down'. This is further complicated by the orientation of the two lone pairs of electrons on the hydroxyl oxygen relative to the lone pair on the amide nitrogen. In all, eight transition states can be identified for 'ω rotation' taking into consideration all positions of the lone pair of electrons on the amide nitrogen and hydroxyl oxygen atoms.

Further, proceeding from the GM and LM structures three TS corresponding to rotation about the τ angle have been identified. Last, there also exists a third type of TS for inversion of the pyramidal state of nitrogen leading to a planar arrangement of the amide nitrogen. In summation, a total of 14 TS have been identified on the potential energy surface of NMAOH (**II**). The relative energies of the minima and TS at the HF, B3LYP, and MP2 levels of theory are listed in Table 1 (The absolute values have been provided in the Supporting Information Table 1A.) The geometries of the minima and TS have been pictorially depicted in Figure 2, and the geometrical data (bond lengths, bond angles, and torsion angles) are given in Table 3.

All structures, except the transition state for inversion of the nitrogen (PyTS), exhibit C_1 symmetry; PyTS has a C_s symmetry. At the HF/6-31+G* level of theory the ranking of the global minimum and local minimum are inverse of that observed at the B3LYP and MP2 levels of theory. It appears that the consideration of electron correlation in both

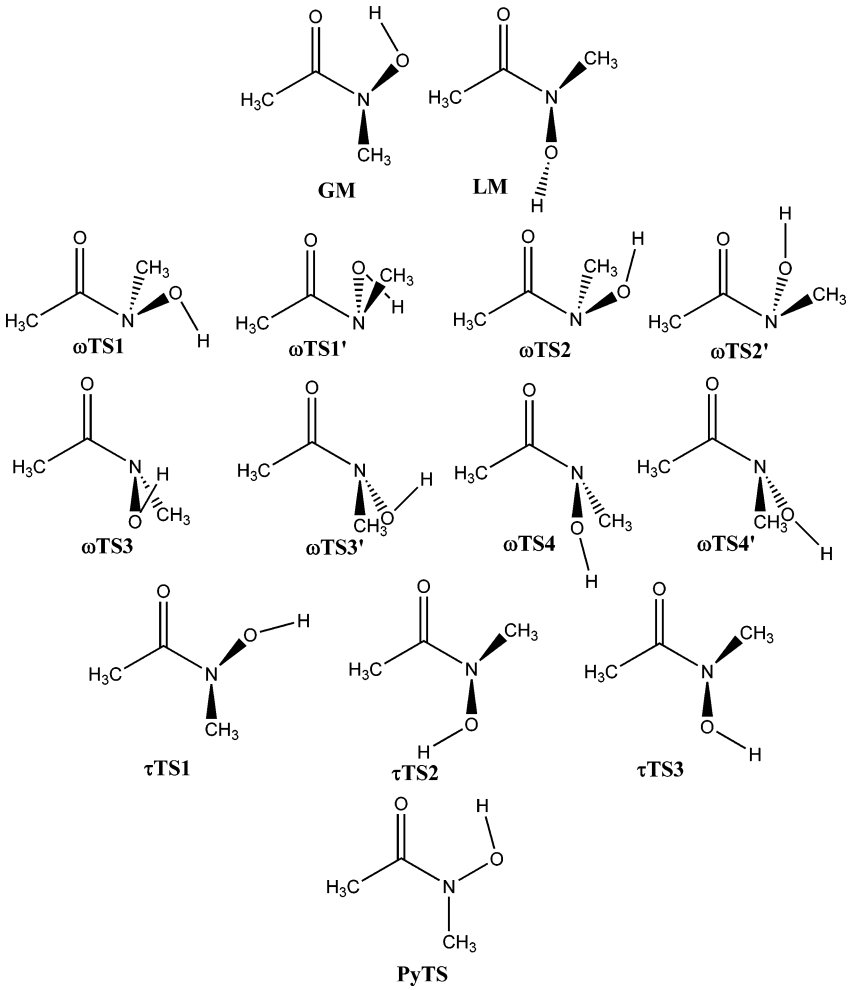


Figure 2. Ground and transition stets of NMAOH.

Table 2. Relative Energies (kcal/mol) of Various Minima and Transition States on PES of BMAOH (**III**) at the HF, B3LYP, and MP2 Levels of Theory Using the 6-31+G* Basis Set[c]

				HF/6-31+G*	B3LYP/6-31+G*	MP2(full)/6-31+G*		
		NIMAG	PG	rel.[a]	rel.[a]	rel.[a]	ω[b]	τ[b]
minima	GM	0	C_s	0.0	0.0	0.0	0	0
	LM	0	C_1	1.9	2.7	2.8	153/−153	180
ω rotation TS	ωTS1	1	C_s	4.3	4.7	4.8	180	0
	ωTS2	1	C_s	6.1	6.5	6.6	0	180
τ rotation TS	τTS	1	C_1	12.5	14.0	15.1	99/−99	86

[a] Relative energy in kcal/mol. [b] Torsion angle in degrees. [c] NIMAG = number of imaginary frequency, PG = point group, GM = global minimum, LM = local minimum.

the B3LYP and MP2 methods resolves this position. The GM is characterized by an intramolecular hydrogen bond between the CO and OH groups forming a five-membered cyclic structure. The O−O distance is 2.566 Å, and the H-bond angle (O−H···O) is 120.4°. Although there is a very small difference in the energies of the GM and LM at the MP2 level, the rotation barrier to interconversion is significant as seen in the energy of the corresponding TS. The rotation barrier to the inversion of nitrogen (PyTS) is negligible (0.4 kcal/mol) at the MP2 level of theory. In the LM the N−O lone pair of electrons is *anti-clinal* exactly like the global minimum of hydroxylamine.

There is a small increase of about 0.08−0.1 Å in the C(O)−N bond length in the transition states for rotation about the ω angle compared to the two ground states (GM and LM). Aubry et al.[28] have reported the crystal structure of a small *N*-hydroxy unnatural peptide ^{t}BuCO-Ψ[CO−N(OH)]-Gly-NH^{i}Pr. The reported structure has a close resemblance to the LM of NMAOH around the *N*-hydoxy amide region. The bond lengths and angles of NMAOH at the MP2(full)/6-31+G* level of theory are close to those in the crystal structure around the *N*-hydroxy amide region (Table 3).

PES of BMAOH (III). The potential energy surface of BMAOH is characterized by two minima—the global minimum (GM) and a local minimum (LM); two transition states for rotation about the ω angle ωTS1 and ωTS2—one arising from the GM and the second from the LM; and one transition state for rotation about the τ angle (τTS). As boron adopts a planar structure (not pyramidal as N in NMAOH), there is no transition state for inversion of boron. The conformations

Table 3. Bond Length (Å) and Bond Angles (deg) of NMAOH (**II**) Optimized at the MP2(full)/6-31+G* Level[a]

parameter	GM	LM	ωTS1	ωTS2	ωTS3	ωTS4	τTS1	τTS2	τTS3	PyTS
CC (1,2)	1.506	1.505	1.498	1.492	1.506	1.512	1.515	1.513	1.504	1.507
CO (2,3)	1.245	1.232 (1.249)	1.218	1.223	1.219	1.217	1.227	1.238	1.235	1.251
CN (2,4)	1.366	1.395 (1.396)	1.475	1.469	1.466	1.462	1.394	1.372	1.387	1.348
NC (4,5)	1.449	1.457 (1.452)	1.466	1.462	1.469	1.455	1.454	1.448	1.449	1.439
NO (4,6)	1.416	1.428 (1.407)	1.464	1.447	1.466	1.442	1.433	1.430	1.441	1.405
OH (6,7)	0.992	0.977	0.977	0.987	0.977	0.986	0.977	0.976	0.977	0.994
CCO (1,2,3)	123.3	123.6	125.4	125.7	123.6	122.7	122.7	122.4	123.9	123.2
CCN (1,2,4)	117.3	116.4	112.5	113.1	118.3	118.5	114.7	116.7	115.9	117.8
CNC (2,4,5)	125.9	118.4	109.5	111.8	114.7	115.1	121.4	121.7	119.6	132.7
CNO (2,4,6)	113.7	112.9	103.5	103.5	101.8	105.6	110.5	120.5	112.3	115.7
CNO (5,4,6)	109.4	110.7	105.7	108.6	105.5	109.3	112.3	108.6	113.9	111.6
NOH (4,6,7)	101.1	103.6	101.4	105.5	101.7	107.1	104.8	106.2	104.5	100.5
CCNC (ω) (1,2,4,5)	31.5	−158.0 (−163.6)	125.5	134.7	36.4	−39.1	38.7	−164.9	−158.6	0.0
CCNO (1,2,4,6)	171.3	−26.3	−122.2	−108.6	−77.0	81.5	173.4	−21.7	−21.1	180.0
CNOH (τ) (2,4,6,7)	9.5	119.4 (119.0)	122.7	−59.1	−105.2	−73.8	−145.9	7.4	−146.7	0.0
OCNC (3,2,4,5)	−151.3	25.4	−54.3	−46.8	−145.7	140.9	−145.7	18.8	25.4	180.0

[a] The values in the parentheses are from the crystal structure of an *N*-hydroxy peptide ^{t}BuCO-Ψ[CO−N(OH)]-Gly-NH^{i}Pr.[28]

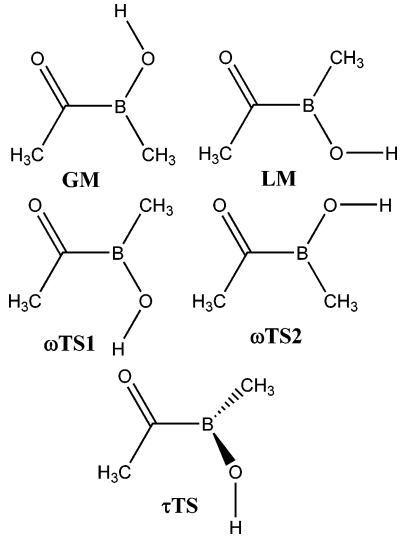


Figure 3. Ground and transition states of BMAOH.

Table 4. Bond Length (Å) and Bond Angles (deg) of BMAOH Optimized at the MP2(full)/6-31+G* Level

parameter	GM	LM	ωTS1	ωTS2	τTS
CC (1,2)	1.503	1.508	1.515	1.513	1.502
CO (2,3)	1.246	1.244	1.241	1.240	1.247
CB (2,4)	1.626	1.608	1.621	1.620	1.593
BC (4,5)	1.564	1.566	1.559	1.575	1.566
BO (4,6)	1.356	1.369	1.369	1.360	1.387
OH (6,7)	0.984	0.974	0.974	0.974	0.968
CCO (1,2,3)	120.8	121.0	120.0	120.2	121.9
CCB (1,2,4)	122.8	120.6	121.1	119.8	126.4
CBC (2,4,5)	124.8	121.1	120.9	122.1	122.0
CBO (2,4,6)	113.6	114.0	119.6	114.1	116.4
CBO (5,4,6)	121.5	124.9	119.5	123.8	121.6
BOH (4,6,7)	108.3	113.3	114.4	113.0	122.6
CCBC (ω) (1,2,4,5)	0.0	−152.7	180.0	0.0	99.1/−99.1
CCBO (1,2,4,6)	180.0	−27.8	0.0	180.0	79.8
CBOH (τ) (2,4,6,7)	0.0	−178.1	0.0	180.0	86.3
OCBC (3,2,4,5)	180.0	−28.9	0.0	180.0	83.3

of the ground and TS of BMAOH are shown in Figure 3, and the relative energies at the HF, B3LYP, and MP2(full) levels of theory with the 6-31+G* basis set are given in Table 2 (The absolute values have been provided in the Supporting Information Table 2A.) The LM and the structure corresponding to the transition state for τ rotation (τTS) exhibit a C_1 symmetry, while the remaining three structures; namely the GM and transition states for ω rotation (ωTS1 and ωTS2) exhibit a C_s symmetry. In the ground-state structures, the ω and τ values in the GM are 0° and 0°, while in the LM they are 150° and 180°, respectively. In ωTS1 and ωTS2, the ω and τ angles have values of 0° and 180° and 180° and 0°, respectively. In the case of τTS, the values of both these torsion angles are 90°.

The geometric parameters of the minima and all transition states of BMAOH at the MP2(full)/6-31+G* level of theory are given in Table 4. In the GM, the −OH group is intramolecularly hydrogen bonded to the CO as is the case with NMAOH; the O−O distance is 2.727 Å, and the H-bond angle (O−H···O) is 117.0°. The hydrogen bond energy, in the case of BMAOH, is roughly estimated as the difference between the LM and GM structures, i.e. ∼2.9 kcal/mol. The changes in the bond lengths from the ground to the TS are relatively small. Some geometric parameters for alkylboranes, arylboranes, and borane complexes have been reported, but there are no experimental data for acylboranes such as BMA[14] and BMAOH. We had earlier reported the geometry of an acylborane BMA, the boron isostere of NMA (**I**), at the QCISD/6-31G* level of theory.[14] The B−O bond length in BMAOH is found to be 1.36 Å (GM) and 1.37 Å (LM) at the MP2(full)/6-31+G* level of theory. The B−O bond length, reported in the literature for a range of organic and inorganic boron containing molecules, is 1.34−1.42 Å (average 1.38 Å) with trigonal planar geometry and 1.39−1.52 Å (average 1.48 Å) for tetrahedral geometry.[29,30] The B−O bond length for BMAOH calculated in this study is in the range of the experimental values. The NBO calculations on the minimum energy structure of BMAOH at the MP2-(full)/6-31+G* level of theory reveal a double bond character for the B−O bond, and the second B−O bond has an

occupancy of 1.99 (~2.0) electrons being contributed from one of the lone pairs of electrons of oxygen. In alkylboranes, the B−C (aliphatic carbon) bond length is about 1.590 Å as in for e.g. dimethylborane,[31] 1.596 Å in dimesityborane,[32] 1.570 Å in ditripylborane,[33] and 1.571 Å in BMA.[14] The B4−C5 bond length in BMAOH is 1.564 Å, which comes near to the experimental value for the aliphatic carbon−boron bond.

Rotation Barrier in NMAOH and BMAOH. The barrier to rotation about the ω angle in the natural peptide is 16.0−25.0 kcal/mol,[34] while that for the boron isostere is about 5.0 kcal/mol.[14] The boron analogues are thus relatively more flexible than the natural peptides. In case of *N*-hydroxy peptides and the corresponding boron isosteres, there are two rotation barriers governed by the ω and τ angle. In the example of NMAOH, the ω rotation barrier is relatively higher (12.6−20.3 kcal/mol) than the τ rotation barrier (6.3−12.2 kcal/mol). In BMAOH, the τ rotation barrier is comparatively higher (15.1 kcal/mol) than the ω rotation barrier (4.8−6.6 kcal/mol). The relative higher τ rotation barrier in boron peptides is a consequence of the B−O double bond character as revealed by NBO calculations.

The rotation barrier in amide systems (like peptides, urea, guanidine, etc.) has been attributed to delocalization of the lone pair of electrons on nitrogen onto the C−N bond as explained by the classical resonance model.[35] This imparts a partial double bond character to the C−N bond. But recent experimental and theoretical studies[36−40] tell a different tale. The electron delocalization in the amide system has been attributed to second-order orbital interactions namely, $n_O \rightarrow \sigma^*_{C-N}$ (delocalization from lone pairs on carbonyl oxygen into the sigma antibonding orbital of the C−N bond i.e. negative hyperconjugation) and $n_N \rightarrow \pi^*_{C-O}$ (delocalization from the lone pair on amide nitrogen to the pi antibonding orbital of the carbonyl group). The energy $E^{(2)}$ associated with negative hyperconjugation i.e. $n_O \rightarrow \sigma^*_{C-N}$ is 29.1 kcal/mol (occupancy of n_O is 1.902 and σ^*_{C-N} is 0.063) and that with $n_N \rightarrow \pi^*_{C-O}$ is 64.1 kcal/mol (occupancy of n_N is 1.772 and π^*_{C-O} is 0.214) for the global minimum of NMAOH at the MP2(full)/6-31+G* level. In case of BMAOH, the energy associated with negative hyperconjugation i.e. $n_O \rightarrow \sigma^*_{C-B}$ is only 11.7 kcal/mol (occupancy of n_O is 1.936 and σ^*_{C-B} is 0.040), indicating that the C−B bond delocalization is insignificant, as a result of which the rotation barrier in boron amides is very small. Thus, in BMAOH, the C−B bond has an essentially single bond character, while the C−N bond in NMAOH has a larger double bond character. The boron peptides are thus far more flexible than the *N*-hydroxy peptides. There is also a $n_O \rightarrow \sigma^*_{O-H}$ interaction i.e. delocalization of the lone pair of electrons on the carbonyl oxygen into the sigma antibonding orbital of the O−H bond which is observed in the global minimum energy structures of both NMAOH and BMAOH but absent in the local minimum structure which affirms the presence of an intramolecular hydrogen bond between CO and OH in the GM of both molecules.

Partial Atomic Charges of NMAOH and BMAOH. The "natural charges" derived from NPA for the global minimum energy structure of NMAOH (**II**) and BMAOH (**III**) are given in Table 5. Replacement of nitrogen by boron decreases the positive charge on the carbonyl oxygen and increases the negative charge on C5 methyl carbon. In BMAOH (**III**), the boron atom has a much greater positive charge than the carbonyl carbon (1.056 vs 0.344). The site for nucleophilic attack in case of NMAOH (**II**) is normally the carbonyl carbon. In BMAOH (**III**), a nucleophile will be drawn toward boron rather than the carbonyl group. This preference for boron as the site for nucleophilic attack is also evident in the 'ESP fitted charges', even though the partial charges differences are of a smaller magnitude. This was the basis of our hypothesis, used to design boron peptides[13] as potential inhibitors of the enzyme serine protease. Figure 4 shows the plausible mechanism by which the boron peptide can act as k_{cat} inhibitor of serine protease. The hydroxyl group of serine in the active site is the nucleophile which attacks the carbonyl carbon of the amide of the substrate peptide, leading to a final hydrolysis of the substrate. When the boron peptide is present in the active site, the hydroxyl group of serine preferentially attacks boron instead of the carbonyl carbon and forms a tetrahedral covalent complex leading to irreversible inhibition of the enzyme. The inhibitors of serine protease could have a potential application in therapeutics.

Table 5. Partial Atomic Charges of NMAOH (**II**) and BMAOH (**III**) Calculated Using NPA and the 'ESP Fit' as per Merz−Singh−Kollman Scheme at the MP2(full)/6-31+G* Level

		natural charges		esp fitted charges	
atom	atom no.	NMAOH (**II**)	BMAOH (**III**)	NMAOH (**II**)	BMAOH (**III**)
C	1	−0.735	−0.737	−0.513	−0.379
C	2	0.818	0.344	0.833	0.484
O	3	−0.753	−0.653	−0.677	−0.580
X	4	−0.279	1.056	−0.263	0.652
C	5	−0.408	−1.059	−0.131	−0.556
O	6	−0.628	−0.972	−0.495	−0.736
H	7	0.542	0.544	0.452	0.426
H	8	0.266	0.249	0.150	0.133
H	9	0.248	0.241	0.164	0.096
H	10	0.246	0.241	0.154	0.096
H	11	0.243	0.254	0.152	0.157
H	12	0.221	0.245	0.065	0.102
H	13	0.217	0.245	0.109	0.102

Conformations of Ala-NOH (IV). The preferred ω and τ angles in NMAOH (**II**) were fixed for Ala-NOH (**IV**), and the (ϕ, ψ) space of Ala-NOH was scrutinized (Table 6). With an *omega* value of 30° and a *tau* value of 10°, the global minimum corresponds to a structure with $\phi = -85°$ and $\psi = -30°$ (Figure 5a). These values are close to the values for a residue at the $i+1$ position in a Type I β-turn ($\phi = -60°$, $\psi = -30°$). The local minimum within 5.0 kcal/mol of the GM has $\phi = 60°$ and $\psi = 50°$ (Figure 5b). These are values of a left-handed alpha helix ($\phi = 57°$, $\psi = 47°$). Both minima display a regular secondary structure motif, which falls in the "allowed regions" of the Ramachandran map. The GM and LM structures are distinguished by two intramolecular hydrogen bonds (Figure 4, parts a and b, respectively)), one between the carbonyl oxygen and the

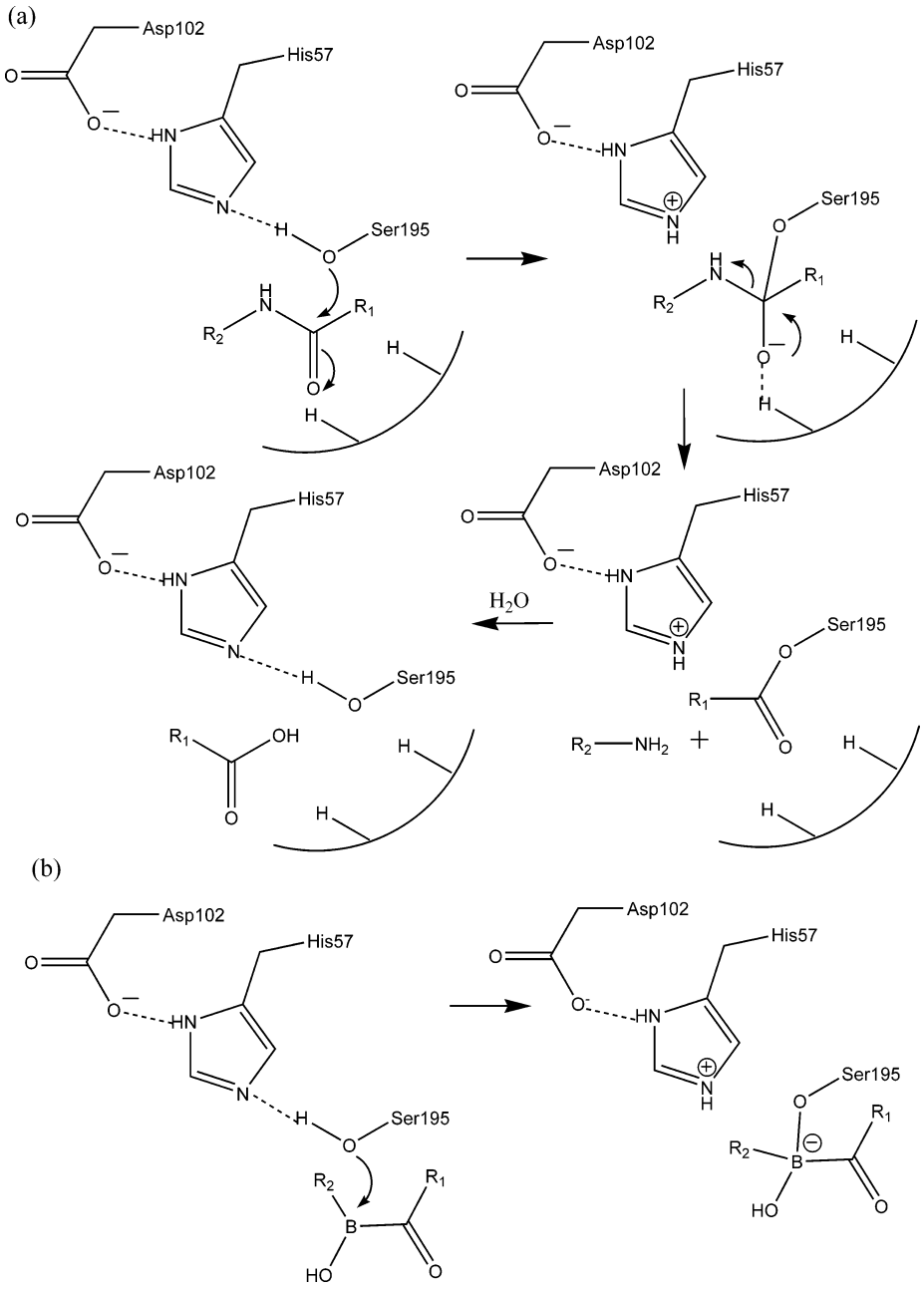


Figure 4. (a) Mechanism of normal substrate hydrolysis by serine protease. (b) Tetrahedral complex of boron peptide with the active site serine.

N-hydroxyl OH, forming a five-membered ring, and the second is found between the *N*-hydroxyl oxygen and the amide NH, figuring a six-membered ring. With an ω angle of 200° and a τ angle of 120°, there is only one favored structure for Ala-NOH (**IV**) with $\phi = -90°$ and $\psi = 140°$. This structure is characterized by only one intramolecular hydrogen bond (Figure 5c) between the *N*-hydoxyl OH and the carbonyl oxygen outlining a six-membered ring. Thus, all the preferred conformations of Ala-NOH are characterized by the presence of one or two intramolecular hydrogen bonds and are conformationally rigid.

Conformations of Ala-BOH (V). In a similar manner, the ϕ, ψ preferences of Ala-BOH (**V**) were investigated, and the results are shown in Table 6. With an *omega* value of 0°, there are two conformations observed within 5.0 kcal/mol of the global minimum energy conformer. The global minimum corresponds to a structure with $\phi = 50°$ and $\psi = -150°$ (Figure 6a), while the local minimum relates to a structure with $\phi = -60°$ and $\psi = 150°$ (Figure 6b). The global minimum shows a strong preference for a positive ϕ value, and ψ in both structures adopts an extended state. The two structures exhibit an intramolecular hydrogen bond (Figure 6a,b) like the one seen in the GM of BMAOH i.e. the preceding carbonyl oxygen and the hydroxyl group on boron are locked, forming a five-membered ring. With an ω angle of 150° and a τ value of 180°, the only structure energetically favored is with $\phi = -160°$ and $\psi = 140°$. The structure is characterized by an intramolecular hydrogen bond (Figure 6c) between the hydroxyl group and the succeeding carbonyl oxygen forming a six-membered ring. Thus, all favored conformations of Ala-BOH exhibit at least one intramolecular hydrogen bond.

Table 6. Conformations and Energies of Ala-NOH (**IV**) and Ala-BOH (**V**)[a]

ω	τ	ϕ	ψ	rel. E (kcal/mol)
		Ala-NOH		
30°	30°	−85°	−30°	0.0
		60°	50°	3.83
200°	120°	−90°	140°	0.0
		Ala-BOH		
0°	0°	50°	−150°	0.0
		−60°	150°	0.13
150°	180°	−160°	140°	0.0

[a] The ω, ϕ, and ψ space of *N*-hydroxy-*N*-methylacetamide and *N*-acetyl-*N*′-hydroxy-*N*′-methylamide of alanine and their boron isosteres.

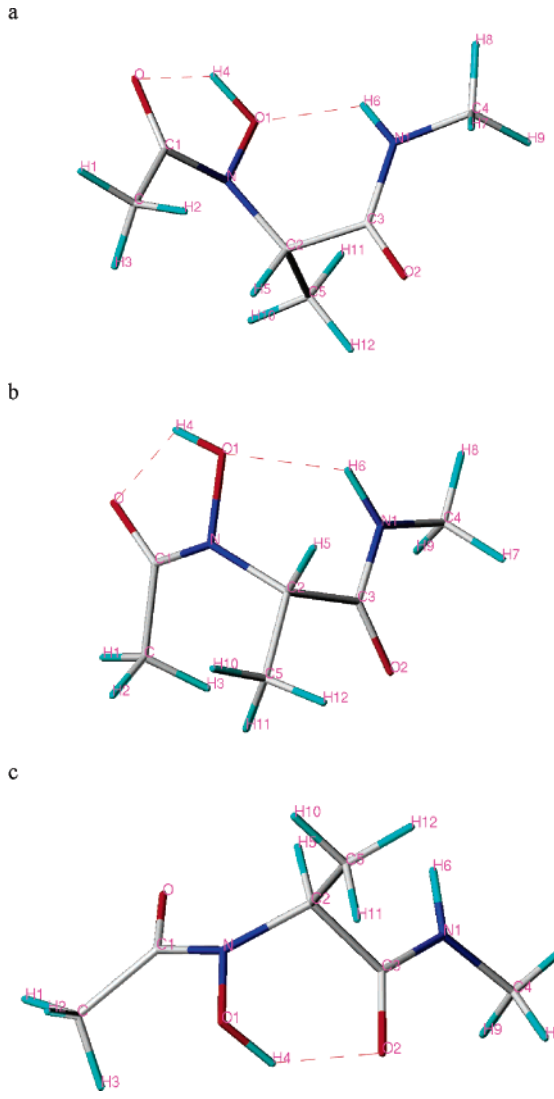


Figure 5. Preferred conformations of Ala-NOH: (a) $\omega = 30°$, $\Phi = -85°$, $\psi = -30°$, (b) $\omega = 30°$, $\Phi = 60°$, $\psi = 50°$, and (c) $\omega = 200°$, $\Phi = -90°$, $\psi = 140°$.

Conclusions

In previous papers[13,14] we had designed a boron isostere of an amino acid by replacement of the amide nitrogen with boron, with the intention of developing an inhibitor of the enzyme serine protease. The synthetic feasibility of such a molecule is a big challenge. As a result, we have modified

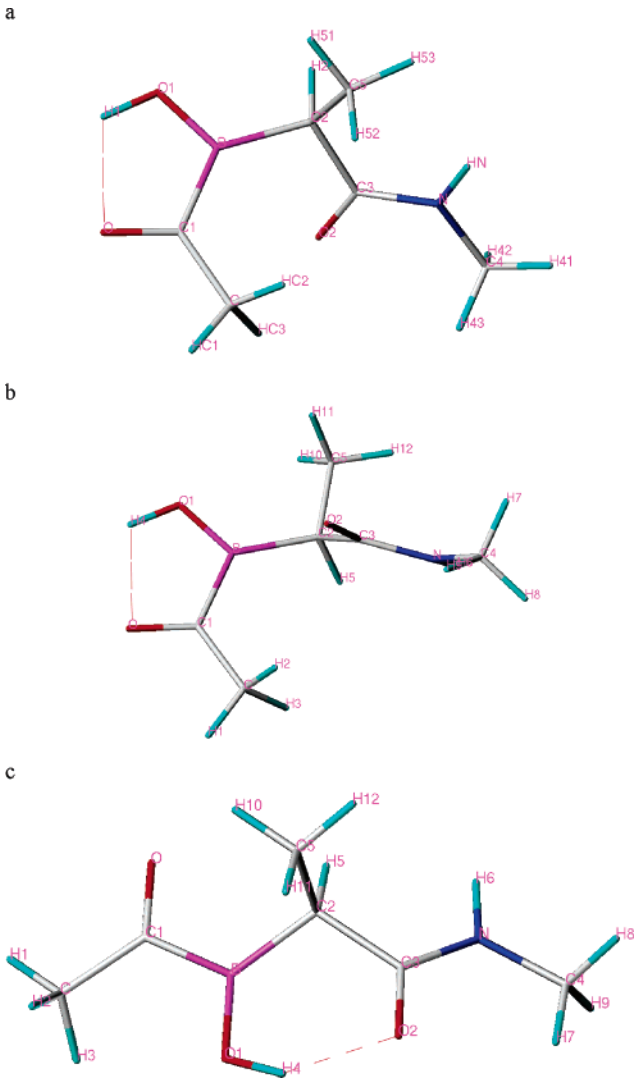


Figure 6. Preferred conformations of Ala-BOH: (a) $\omega = 0°$, $\Phi = 50°$, $\psi = -150°$, (b) $\omega = 0°$, $\Phi = -60°$, $\psi = 150°$, and (c) $\omega = 150°$, $\Phi = -160°$, $\psi = 140°$.

the molecule by replacing B−H with B−OH, which should make it relatively easy to synthesize; and this also has an analogy with *N*-hydroxy amides which are well-known. The conformational space of *N*-hydroxy peptides and their boron isosteres has been the focus of investigation in this paper. The minimum in the *omega* torsion space of such molecules has been identified using *N*-hydroxy-*N*-methylacetamide (NMAOH) and acetylmethylhydroxyborane (BMAOH) as model peptides. The ground and various transition states have been calculated at the HF, B3LYP, and MP2(full) levels of theory with the 6-31+G* basis set. The *omega* rotation barrier is 12.6−20.3 kcal/mol for the *N*-hydroxy peptide (NMAOH) and 4.8−6.6 kcal/mol for its corresponding boron isostere, BMAOH. The difference in the rotation barriers has been attributed to second-order orbital interactions, mainly negative hyperconjugation. The global minimum energy conformation of both molecules exhibits an intramolecular hydrogen bond between the carbonyl oxygen and the hydroxyl group which confers some rigidity to the conformation. The barrier for rotation about the torsion angle τ i.e. rotation about N−O and B−O bonds is 6.3−12.2 kcal/mol

for the *N*-hydroxy peptide and is 15.1 kcal/mol for the boron isostere. The elevated value for the boron isostere has been attributed to the single bond character of the N−O bond against the double bond character of the B−O bond. The replacement of nitrogen by boron also significantly changes the charge distribution in these molecules. A relatively greater positive charge on the boron atom over the carbonyl carbon makes boron the preferential site of attack by a nucleophile in boron peptides, which otherwise occurs on the carbonyl carbon in the natural peptides. This observation can be potentially exploited for the design of serine protease inhibitors. It would be interesting to study the transition state barrier for hydrolysis at the carbonyl carbon versus the boron, which is the next step in the study. The minimum energy structures of NMAOH and BMAOH were then used to study the ϕ and ψ preferences in *N*-acetyl-*N*′-hydroxy-*N*′-methylamide of alanine (Ala-NOH) and its boron isostere (Ala-BOH). Ala-NOH demonstrates conformations with Type-I beta turn, left-handed α-helix, positive ϕ values and extended ψ states. Ala-BOH, on the other hand, favors conformations with positive ϕ and extended ψ values. In previous work on natural peptides and their boron isosteres, we had noticed a much lower barrier to rotation about the ω angle and a unique preference for positive *phi* values in the boron analogues. The boron isosteres of *N*-hydroxy peptide also show a similar tendency. In conclusion, the replacement of nitrogen by boron in natural and N-hydroxy peptides causes a significant change in the conformational space and electronic properties, and these features can be profitably exploited to design peptides with specific geometries and chemical attributes.

Acknowledgment. This work was supported by the Department of Science and Technology, New Delhi through their FIST program (SR/FST/LS1-163/2003). A.K.M. thanks the University Grants Commission, New Delhi and the Council of Scientific and Industrial Research, New Delhi, and S.A.K. thanks the Lady Tata Memorial Trust, Mumbai for financial support.

Supporting Information Available: Absolute energy values (au) of conformations of NMAOH (**II**) and BMAOH (**III**) (Tables 1A and 2A, respectively) and XYZ coordinates of the global minimum of structures **II** and **III** optimized at the MP2(full)/6-31+G* level of theory and of structures **IV** and **V** optimized at the B3LYP/6-31+G* level of theory (Tables 7−10). This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Ramachandran, G. N.; Sasisekharan, V. Conformation of Polypeptides and Proteins. *Adv. Protein Chem.* **1968**, *28*, 283−437.

(2) Hagler, T. A.; Leiserowitz, L.; Tuval, M. Experimental and Theoretical Studies of the Barrier to Rotation about N−Cα and Cα-C′ Bonds (ϕ and ψ) in Amides and Peptides. *J. Am. Chem. Soc.* **1976**, *98*, 4600−4612.

(3) Hruby, V. J. Designing Peptide Receptor Agonists and Antagonists. *Nat. Rev. Drug Discovery* **2002**, *1*, 847−858.

(4) Vogen, S. M.; Paczkowski, N. J.; Kirnarsky, L.; Short, A.; Whitmore, J. B.; Sherman, S. A.; Taylor, S. M.; Sanderson, S. D. Differential Activities of Decapeptide Agonists of Human C5a: The Conformational Effects of Backbone *N*-Methylation. *Int. Immunopharmacol.* **2001**, *12*, 2151−62.

(5) Ye, Y.; Liu, M.; Kao, J. L.; Marshall, G. R. Peptide-bond Modification for Metal Coordination: Peptides Containing Two Hydroxamate Groups. *Biopolymers* **2003**, *71*, 489−515.

(6) Fischer, P. M. The design, synthesis and application of stereochemical and directional peptide isomers: a critical review. *Curr. Protein Pept. Sci.* **2003**, *4*, 339−356.

(7) Baldauf, C.; Günther, R.; Hofmann, H. J. Conformational Properties of Sulphonamido Peptides. *J. Mol. Struct. (THEOCHEM)* **2004**, *675*, 19−28.

(8) Kettner, C. A.; Shenvi, A. B. Inhibition of Serine Protease Leukocyte Elastase, Pancreatic Elastase, Cathepsin G, and Chymotrypsin by Peptide Boronic Acids. *J. Biol. Chem.* **1984**, *259*, 15106−15114.

(9) Kettner, C. A.; Bone, R.; Agard, D. A.; Bachovchin, W. W. Kinetic Properties of the Binding of α-Lytic Protease to Peptide Boronic Acids. *Biochemistry* **1988**, *27*, 7682−7688.

(10) Spielvogel, B. F.; Wojnowich, L.; Das, M. K.; McPhail, A. T.; Hargrave, K. D. Boron Analogues of Amino Acids. Synthesis and Biological Activity of Boron Analogues of Betaine. *J. Am. Chem. Soc.* **1976**, *98*, 5702−5703.

(11) Spielvogel, B. F.; Das, M. K.; McPhail, A. T.; Onam, K. D.; Hall, I. H. Boron Analogues of the alpha-Amino Acids. Synthesis, X-Ray Crystal Structure, and Biological Activity of Ammonia-Carboxyborane, the Boron Analogue of Glycine. *J. Am. Chem. Soc.* **1980**, *102*, 6343−6344.

(12) Miller, M. C.; Sood, A.; Spielvogel, B. F.; Hall, I. H. Synthesis and Antitumor Activity of Boronated Dipeptides containing Aromatic Amino Acids. *Anticancer Res.* **1997**, *5A*, 3299−3306.

(13) Datar, P. A.; Coutinho, E. C. The ϕ, ψ Space of Boron Isosteres of Amino Acids: An *Ab Initio* Study. *J. Theor. Comput. Chem.* **2004**, *3*, 189−202.

(14) Malde, A. K.; Khedkar, S. A.; Coutinho, E. C.; Saran A. Geometry, Transition States, and Vibrational Spectra of Boron Isostere of *N*-methylacetamide by *Ab Initio* Calculations. *Int. J. Quantum Chem.* **2005**, *102*, 734−742.

(15) Hehre, W. J.; Random, L.; Schleyer, P. V. R.; Pople, J. A. In *Ab Initio Molecular Orbital Theory;* Wiley: New York, 1985.

(16) Parr, R. G.; Yang, W. In *Density Functional Theory of Atoms and Molecules;* O.U.P.: New York, 1989.

(17) *Gaussian 03, Revision C.01*; 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.; 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.: Wallingford, CT, 2004.

(18) Roothan, C. C. New Developments in Molecular Orbital Theory. *Rev. Mod. Phys.* **1951**, *23*, 69−89.

(19) Becke, A. D. Density functional thermochemistry. III. The role of exact exchange. *J. Chem. Phys.* **1993**, *98*, 5648−5652.

(20) Lee, C.; Yang, W.; Parr, R. G. Development of the Colle-Salvetti Correlation-Energy Formula into a Functional of the Electron Density. *Phys. Rev.* **1988**, *37B*, 785−789.

(21) Perdew, J. P.; Wang, Y. Accurate and simple analytic representation of the electron-gas correlation energy. *Phys. Rev.* **1992**, *45B*, 13244−13249.

(22) Møller, C.; Plesset, M. S. Note on an Approximation Treatment for Many-Electron Systems. *Phys. Rev.* **1934**, *46*, 618−622.

(23) Martin-Head, G.; Pople, J. A.; Frisch, M. J. MP2 Energy Evaluation by Direct Methods. *Chem. Phys. Lett.* **1988**, *153*, 503−506.

(24) Glendening, E. D.; Reed, A. E.; Carpenter, J. E.; Weinhold, F. NBO Version 3.1.

(25) Reed, A. E.; Weinstock, R. B.; Weinhold, F. Natural Population Analysis. *J. Chem. Phys.* **1985**, *83*, 735−746.

(26) Reed, A. E.; Weinhold, F.; Curtiss, L. A. Intermolecular interactions from a natural bond orbital, donor−acceptor viewpoint. *Chem. Rev.* **1988**, *88*, 899−926.

(27) Singh, U. C.; Kollman, P. A. An Approach to Computing Electrostatic Charges for Molecules. *J. Comput. Chem.* **1984**, *5*, 129−145.

(28) Aubry, A.; Dupont, V.; Marraud, M. ^{t}BuCO-Ψ[CO−N(OH)]-Gly-NH^{i}Pr. *Acta Crystallogr., Sect. C*: *Cryst. Struct. Commun*. **1995**, *51*, 1577−1579.

(29) Filatov, S.; Shepelev, Y.; Bubnova, R.; Sennova, N.; Egorysheva, A. V.; Kargin, Y. F. The study of $Bi_3B_5O_{12}$: synthesis, crystal structure and thermal expansion of oxoborate $Bi_3B_5O_{12}$. *J. Solid State Chem.* **2004**, *177*, 515−522.

(30) Shishkov, I. F.; Khristenko, L. V.; Rudakov, F. M.; Vilkov, L. V.; Karlov, S. S.; Zaitseva, G. S.; Samdal, S. The molecular structure of boratrane determined by gas electron diffraction and quantum mechanical calculations. *J. Mol. Struct.* **2002**, *641*, 199−205.

(31) Vijay, A.; Sathyanarayana, D. N. Effects of Basis Set and Electron Correlation on the Structure and Vibrational Spectra of Diborane. *J. Mol. Struct.* **1995**, *351*, 215−229.

(32) Entwistle, C. D.; Marder, T. B.; Smith, P. S.; Howard, A. K.; Fox, M. A.; Mason, S. A. Dimesitylborane monomer−dimer equilibrium in solution, and the solid-state structure of the dimer by single-crystal neutron and X-ray diffraction. *J. Organomet. Chem.* **2003**, *680*, 165−172.

(33) Bartlett, R. A.; Rasikadis, H. V.; Olmstead, M. M.; Power, P. P.; Weese, K. J. Synthesis of the Monomeric $HBtrip_2$ (trip − 2,4,6-iso-$Pr_3C_6H_2$) and the X-ray Crystal Structures of $[HBMes_2]_2$ (Mes = 2, 4, 6,-$Me_3C_6H_2$) and $HBtrip_2$. *Organometallics* **1990**, *9*, 146−150.

(34) Villani, V.; Alagona, G.; Ghio, C. Ab Initio Studies on *N*-Methylacetamide. *Mol. Eng.* **1999**, *8*, 135−153.

(35) Pauling, L. *In The Nature of Chemical Bond*; Cornell University Press: Ithaca, 1960.

(36) Bharatam, P. V.; Iqbal, P.; Malde, A.; Tiwari, R. Electron Delocalization in Aminoguanidines: A Computational Study. *J. Phys. Chem.* **2004**, *108*, 10509−10517.

(37) Bharatam, P. V.; Moudgil, R.; Kaur, D. Electron Delocalization in Isocyanates, Formamides, and Ureas: Importance of Orbital Interactions. *J. Phys. Chem.* **2003**, *107*, 1627−1634.

(38) Glendening, E. D.; Hrabal, J. A., II. Resonance in Formamide and Its Chalcogen Replacement Analogues: A Natural Population Analysis/Natural Resonance Theory Viewpoint. *J. Am. Chem. Soc.* **1997**, *119*, 12940−12946.

(39) Lauvergnat, D.; Hiberty, P. C. Role of Conjugation in the Stabilities and Rotational Barriers of Formamide and Thioformamide. An ab initio Valence-Bond Study. *J. Am. Chem. Soc.* **1997**, *119*, 9478−9482.

(40) Wiberg, K. B.; Rush, D. J. Solvent Effects on the Thioamide Rotational Barrier: An Experimental and Theoretical Study. *J. Am. Chem. Soc.* **2001**, *123*, 2038−2046.

CT050242V

JCTC Journal of Chemical Theory and Computation

Oxidative Addition of the Chloromethane C−Cl Bond to Pd, an ab Initio Benchmark and DFT Validation Study

G. Theodoor de Jong and F. Matthias Bickelhaupt*

Afdeling Theoretische Chemie, Scheikundig Laboratorium der Vrije Universiteit, De Boelelaan 1083, NL-1081 HV Amsterdam, The Netherlands

Received October 14, 2005

Abstract: We have computed a state-of-the-art benchmark potential energy surface (PES) for the archetypal oxidative addition of the chloromethane C−Cl bond to the palladium atom and have used this to evaluate the performance of 26 popular density functionals, covering LDA, GGA, meta-GGA, and hybrid density functionals, for describing this reaction. The ab initio benchmark is obtained by exploring the PES using a hierarchical series of ab initio methods [HF, MP2, CCSD, and CCSD(T)] in combination with a hierarchical series of seven Gaussian-type basis sets, up to g polarization. Relativistic effects are taken into account through a full four-component all-electron approach. Our best estimate of kinetic and thermodynamic parameters is −11.2 (−10.8) kcal/mol for the formation of the most stable reactant complex, 3.8 (2.7) kcal/mol for the activation energy of direct oxidative insertion (OxIn), and −28.0 (−28.8) kcal/mol for the reaction energy (all energies relative to separate reactants, zero-point vibrational energy-corrected values in parentheses). Our work highlights the importance of sufficient higher angular momentum polarization functions for correctly describing metal-d-electron correlation. The best overall agreement with our ab initio benchmark is obtained by functionals from all three categories, GGA, meta-GGA, and hybrid DFT, with mean absolute errors of 0.8−3.0 kcal/mol and errors in activation energies for OxIn ranging from 0.0 to 1.2 kcal/mol. For example, three well-known functionals, BLYP, OLYP, and B3LYP, compare very reasonably with, respectively, an underestimation of the barrier for OxIn of −4.2 kcal/mol and overestimations of 4.2 and 1.6 kcal/mol. Interestingly, all important features of the CCSD(T) benchmark potential energy surfaces for the Pd-induced activation of C−H, C−C, C−F, and C−Cl bonds are reproduced correctly within a few kcal/mol by BLYP, OLYP, and B3LYP, while at the same time, none of these functionals is the "best one" in each individual case. This follows from an overall comparison of the results of the present as well as previous studies.

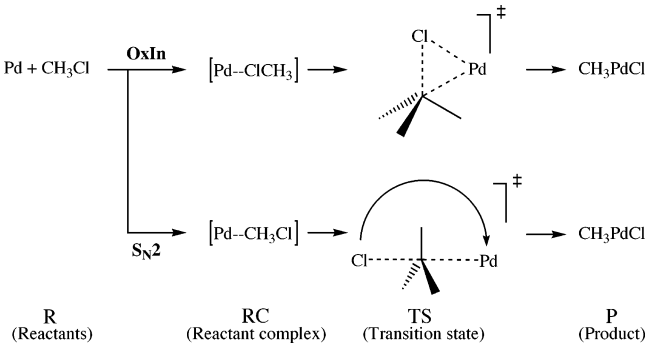
1. Introduction

The catalytic activation of the C−Cl bond is an efficient tool for selectively converting simple educts, via C−C bond formation, into more complex compounds. This process, which is often based on catalytically active palladium complexes, is therefore of major importance for synthetic chemistry. The most intensively used substrates for such C−C coupling reactions are aryl halides, whereas it is more difficult in this context to exploit alkyl halides.[1] While C−H and C−C bond activations have been the subject in various computational investigations, the oxidative addition of C−Cl or, more generally, C−halogen bonds has received less attention.[2] Still, there are a number of computational studies[2−11] on the activation of C−X bonds by d^{10} metal centers, such as palladium complexes, which is one of our main subjects of interest because of its relevance for homogeneous catalysis.[12]

* Corresponding author fax: +31-20-59 87629; e-mail: fm.bickelhaupt@few.vu.nl.

10.1021/ct050254g CCC: $33.50
Published on Web 01/21/2006

Chart 1. Model Reactions and Nomenclature



Transition-metal-induced C−Cl bond activation usually proceeds via an oxidative addition process in which the metal increases its formal oxidation state by two units. There has been controversy about the mechanism of this reaction.[13] One mechanism that has been proposed requires the concerted transfer of two electrons and involves either a concerted front-side displacement or a concerted nucleophilic displacement (S_N2) proceeding via backside attack of the C−Cl bond by the metal. Theoretical studies on the oxidative addition of the C−Cl bond in chloromethane to the Pd atom show that this process can indeed proceed via direct oxidative insertion of the metal into the C−Cl bond (OxIn) or via S_N2 substitution followed, in a concerted manner, by leaving-group rearrangement (S_N2-ra).[3,10] The reaction barrier for OxIn is lower than that for the S_N2 pathway. Interestingly, anion assistance, for example, coordination of a chloride anion to Pd, reverses this order in activation energies and makes S_N2 the preferred pathway. Note that this shift in mechanism also corresponds to a change in stereochemistry at the carbon atom involved, namely, from retention (OxIn) to inversion of configuration (S_N2). This is of practical relevance for substrates in which the carbon atom, C*, is asymmetric (which is obviously not the case in the simple model substrate chloromethane). The two pathways are schematically summarized in Chart 1.

The purpose of the present study is two-fold. In the first place, we wish to obtain a reliable benchmark for the potential energy surface (PES) for the oxidative addition of the C−Cl bond of chloromethane to Pd(0). This is done by exploring this PES with a hierarchical series of ab initio methods {Hartree−Fock (HF), second-order Møller−Plesset perturbation theory (MP2),[14] and coupled cluster theory[15] with single and double excitations (CCSD)[16] and with triple excitations treated perturbatively [CCSD(T)][17]} in combination with a hierarchical series of Gaussian-type basis sets of increasing flexibility and polarization (up to g functions). The basis set superposition error (BSSE) is accounted for by counterpoise correction (CPC).[18] Relativistic effects are treated with a full four-component all-electron approach. To our knowledge, these are the first benchmarking calculations at an advanced correlated level for this model reaction.

The second purpose of our work is to evaluate the performance of 26 popular density functionals, covering LDA, GGA, meta-GGA, and hybrid density functionals, for describing the oxidative addition of the chloromethane C−Cl bond to Pd, using the ab initio benchmark as a reference point. Here, we anticipate that, while the latter turns out to be satisfactory in terms of accuracy and reliability, it is prohibitively expensive if one wishes to study more realistic model catalysts and substrates. Thus, our survey of 26 density functionals serves to validate one or more of these density functional theory (DFT) approaches as a computationally more efficient alternative to high-level ab initio theory in future investigations in the field of computational catalysis.[11] A general concern associated with the application of DFT to the investigation of chemical reactions is its notorious tendency to underestimate activation energies.[6,19−24] However, very recently, with the same approach as has been used in the present study, we investigated the insertion of the Pd d^{10} atom into the C−H bond of methane, the C−C bond of ethane, and the C−F bond of fluoromethane as important archetypal examples of oxidative addition reactions:[25−28] DFT[29−31] turned out to reproduce the highest level ab initio (coupled-cluster) benchmark PESs within a few kilocalories per mole.[26−28] Interestingly, in the case of palladium-induced C−H and C−C bond activation,[26,27] the well-known BLYP functional turned out to be among the best performing functionals, providing PESs that are better than those of most of the high-level meta-GGA and hybrid functionals. On the other hand, the activation of the C−F bond turns out to be somewhat better described by OLYP and B3LYP.[28] Here, we are interested in how far the same conclusions hold for palladium-induced C−Cl bond activation. In addition to evaluating and ranking the performance of the density functionals, we investigate the dependence of the resulting PES on the basis-set size and on the use of the frozen-core approximation. We conclude with a critical overview and comparison of the palladium-induced activations of all bonds for which we have so far carried out an ab initio benchmark and DFT validation study: C−H, C−C, C−F, and C−Cl.

2. Method and Computational Details

2.1. Geometries. All geometry optimizations have been done with DFT using the Amsterdam Density Functional (ADF) program.[32−35] For eight different LDA and GGA functionals, the performances for computing the geometries and relative energies of the stationary points along the PES of our model reaction (see Chart 1) were compared. These density functionals are the LDA functional VWN[36] and the GGA functionals BP86,[37,38] BLYP,[37,39] PW91,[40−43] PBE,[44,45] revPBE,[46] RPBE,[47] and OLYP.[39,48] They were used in combination with the TZ2P basis set, which is a large uncontracted set of Slater-type orbitals (STOs) containing diffuse functions, which is of triple-ζ quality and has been augmented with two sets of polarization functions: 2p and 3d on H, 3d and 4f on C and Cl, 5p and 4f on Pd. The core shells of carbon (1s), chlorine (1s2s2p), and palladium (1s2s2p3s3p3d) were treated by the frozen-core approximation.[32] An auxiliary set of s, p, d, f, and g STOs was used to fit the molecular density and to represent the Coulomb and exchange potentials accurately in each SCF cycle.[32] Relativistic effects were accounted for using the zeroth-order regular approximation (ZORA).[49] For each of the eight functionals, all stationary points were confirmed to be equilibrium structures (no imaginary frequencies) or a

Table 1. Basis Sets Used in the ab Initio Calculations

name	Pd	C	H	Cl
BS1	(24s16p13d)[a]	cc-aug-pVDZ[b]	cc-aug-pVDZ[b]	cc-aug-pVTZ[b]
BS2	(24s16p13d)[a] + 1f	cc-aug-pVDZ[b]	cc-aug-pVDZ[b]	cc-aug-pVTZ[b]
BS2(−)	(24s16p13d)[a] + 1f	cc-aug-pVDZ[b]	cc-aug-pVDZ[b]	cc-aug-pVDZ[b]
BS2(+)	(24s16p13d)[a] + 1f	cc-aug-pVTZ[b]	cc-aug-pVTZ[b]	cc-aug-pVTZ[b]
BS3	(24s16p13d)[a] + 4f	cc-aug-pVDZ[b]	cc-aug-pVDZ[b]	cc-aug-pVTZ[b]
BS4	(24s16p13d)[a] + 4f + p	cc-aug-pVDZ[b]	cc-aug-pVDZ[b]	cc-aug-pVTZ[b]
BS5	(24s16p13d)[a] + 4f + p + g	cc-aug-pVDZ[b]	cc-aug-pVDZ[b]	cc-aug-pVTZ[b]

[a] TZP quality. [b] Completely uncontracted.

transition state (one imaginary frequency) through vibrational analysis. Enthalpies at 298.15 K and 1 atm were calculated from 0 K electronic energies according to the following equation, assuming an ideal gas:

$$\Delta H_{298} = \Delta E + \Delta E_{\mathrm{trans},298} + \Delta E_{\mathrm{rot},298} + \Delta E_{\mathrm{vib},0} + \Delta(\Delta E_{\mathrm{vib},0})_{298} + \Delta(pV)$$

Here, $\Delta E_{\mathrm{trans},298}$, $\Delta E_{\mathrm{rot},298}$, and $\Delta E_{\mathrm{vib},0}$ are the differences between products and reactants in translational, rotational, and zero-point vibrational energies, respectively; $\Delta(\Delta E_{\mathrm{vib},0})_{298}$ is the change in the vibrational energy difference going from 0 to 298.15 K. The vibrational energy corrections are based on our frequency calculations. The molar work term $\Delta(pV)$ is $(\Delta n)RT$; $\Delta n = -1$ for two reactants (Pd + CH_3Cl) combining to one species. Thermal corrections for the electronic energy are neglected.

2.2. Ab Initio Calculations. On the basis of the ZORA-BLYP/TZ2P geometries, energies of the stationary points were computed in a series of single-point calculations with the program package DIRAC[50,51] using the following hierarchy of quantum chemical methods: HF, MP2, CCSD, and CCSD(T). Relativistic effects are accounted for using a full all-electron four-component Dirac−Coulomb approach with a spin-free Hamiltonian.[52] The two-electron integrals exclusively over the small components have been neglected and corrected with a simple Coulombic correction, which has been shown to be reliable.[53]

A hierarchical series of Gaussian-type basis sets was used (see Table 1). For carbon, hydrogen, and chlorine, Dunning's correlation consistent augmented double-ζ (cc-aug-pVDZ) and triple-ζ (cc-aug-pVTZ) basis sets were used.[54,55] These were used in uncontracted form because it is technically difficult to use contracted basis sets in the kinetic balance procedure in DIRAC.[56] The basis set of palladium is based on an uncontracted basis set (24s16p13d), which is of triple-ζ quality, and has been developed by K. Faegri, Jr. (personal communication). The combination of this basis set for palladium and the aforementioned cc-aug-pVDZ basis set for carbon and hydrogen and cc-aug-pVTZ basis set for chlorine is denoted BS1 (see Table 1). As a first extension, in BS2, one set of 4f polarization functions was added with an exponent of 1.472, as reported by Ehlers et al.[57] In BS3, this single set of 4f functions was substituted by four sets of 4f polarization functions as reported by Langhoff and coworkers with exponents of 3.611 217, 1.295 41, 0.554 71, and 0.237 53.[58] Thereafter, going to BS4, an additional set of diffuse p functions was introduced with an exponent of 0.141 196, as proposed by Osanai et al.[59] BS5 was created by adding a set of g functions, with an exponent of 1.031 690 071. This value is close but not exactly equal to the exponent of the g functions optimized by Osanai. Instead, it is equal to the value of one of the exponents of the d set of Faegri, which reduces computational costs.

Note that the basis sets BS1−BS5 used in the present study (see Table 1) correspond in quality to the basis sets BS1−BS5 used in our recent studies on the oxidative addition of Pd to the C−C bond of ethane and the C−F bond of fluoromethane (see Table 2 in ref 27 and Table 1 in ref 28, respectively), using, however, an uncontracted cc-aug-pVTZ basis set for chlorine. For the C−C addition reaction, concerning the uncontracted cc-aug-pVDZ basis set for C and H, relative energies were converged to within ca. 1 kcal/mol at BS5. For the present reaction, we have investigated more extensively how well the relative energies are converged with respect to the basis-set sizes of carbon, hydrogen, and chlorine. To this end, basis sets BS2(−) and BS2(+) were also constructed. BS2(−) is equal to BS2, but with a cc-aug-pVDZ instead of a cc-aug-pVTZ basis set for chlorine. BS2(+) also corresponds to BS2, but with a cc-aug-pVTZ basis set for all three elements C, H, and Cl. For a schematic overview, see Table 1.

2.3. DFT Calculations. On the basis of the ZORA-BLYP/TZ2P geometries, we have also evaluated, in a series of single-point calculations, how the ZORA-BLYP relative energies of stationary points along the PES depend on the basis-set size for four different all-electron (i.e., no frozen-core approximation) STO basis sets, namely, ae-DZ, ae-TZP, ae-TZ2P, and ae-QZ4P, and on the use of the frozen-core approximation. The ae-DZ basis set is of double-ζ quality and is unpolarized for C, Cl, and H but has been augmented with a set of 5p polarization functions for Pd. The ae-TZP basis set is of triple-ζ quality and has been augmented with one set of polarization functions on every atom: 2p on H, 3d on C and Cl, and 5p on Pd. The ae-TZ2P basis set (the all-electron counterpart corresponding to the above-mentioned TZ2P basis that is used in conjunction with the frozen-core approximation) is also of triple-ζ quality and has been augmented with two sets of polarization functions on each atom: 2p and 3d on H, 3d and 4f on C and Cl, and 5p and 4f on Pd. The ae-QZ4P basis set is of quadruple-ζ quality and has been augmented with four sets of polarization functions on each atom (five for chlorine): two 2p and two 3d sets on H, two 3d and two 4f sets on C, three 3d and two 4f sets on Cl, and two 5p and two 4f sets on Pd.

Finally, again on the basis of the ZORA-BLYP/TZ2P geometries, we have computed, in a post-self-consistent-field

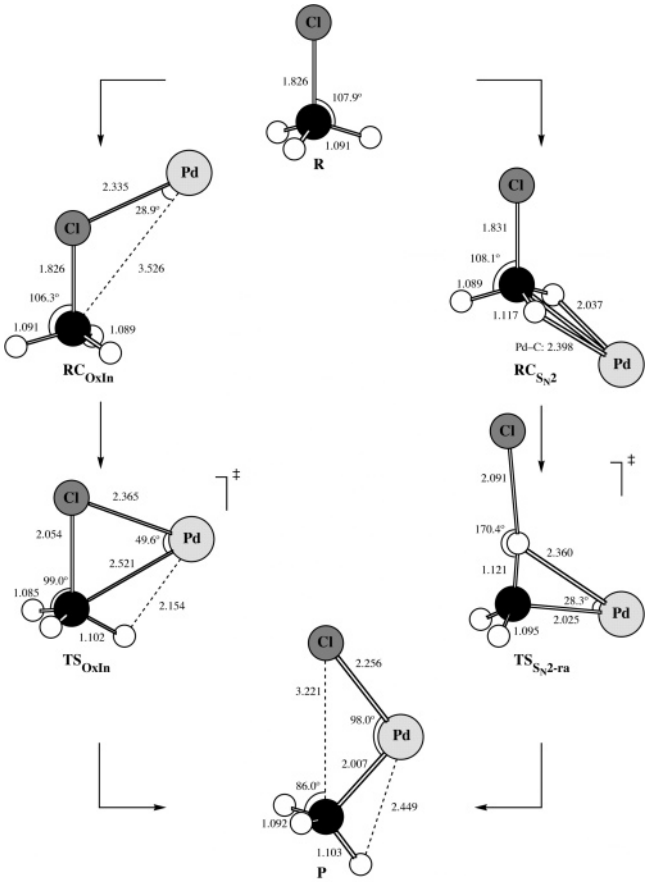


Figure 1. Structures of stationary points along the reaction coordinates of the OxIn- and S_N2-type pathways for oxidative addition of the C−Cl bond of CH_3Cl to Pd. Geometry optimized at ZORA-BLYP/TZ2P, i.e., with frozen-core approximation.

(post-SCF) manner, that is, using in all cases the electron density obtained at ZORA-BLYP/ae-TZ2P, the relative energies of stationary points along the PES for various LDA, GGA, meta-GGA, and hybrid functionals. In addition to the ones used in the geometry optimizations (see Section 2.1), the following density functionals were examined: the GGA functionals Becke 88x + BR89c,[60,61] FT97,[62] HCTH/93,[63] BOP,[60,64] HCTH/120,[65] HCTH/147,[65] and HCTH/407;[66] the meta-GGA functionals BLAP3,[67] VS98,[68] KCIS,[69] PKZB,[70,71] Bmτ1,[72] OLAP3,[48,67] and TPSS;[73,74] and the hybrid functionals B3LYP,[75,76] O3LYP,[77] X3LYP[78] (all based on VWN5[79]), and TPSSh.[73,74]

3. Results and Discussion

3.1. Geometries of Stationary Points and Characteristics of the Addition Reaction. First, we examine the geometries of stationary points along the reaction coordinates of the two pathways for the oxidative addition of Pd to the C−Cl bond of chloromethane, computed with the LDA functional VWN and the GGA functionals BP86, BLYP, PW91, PBE, revPBE, RPBE, and OLYP in combination with the TZ2P basis set, the frozen-core approximation, and the ZORA to account for relativistic effects. For the BLYP functional, the results are given in Figure 1. For the other functionals, the optimized geometries are given in the Supporting Information, in Figure S1 and Table S1.

For each of the functionals, the reaction characteristics are similar. For the direct insertion (OxIn) pathway, the reaction proceeds from the reactants R via the formation of a stable, C_s symmetric reactant complex, RC_{OxIn}, in which the chlorine atom coordinates to Pd, to a transition state, TS_{OxIn}, of C_s symmetry and, finally, a stable product P of C_s symmetry (see Figure 1). For the alternative S_N2 pathway, the reaction proceeds from the reactants via formation of another stable, C_s symmetric reactant complex, RC_{S_N2}, in which chloromethane coordinates via two hydrogen atoms in an η^2 fashion to Pd (see Figure 1), completely analogous to reactant complexes for the reaction of Pd with methane,[26] ethane,[27] and fluoromethane.[28] From RC_{S_N2}, the S_N2 substitution then occurs in concert with a rearrangement of the Cl^- leaving group from carbon to palladium, with a transition state TS_{S_N2-ra} of C_s symmetry and, finally, the same product P as in the OxIn pathway.

We wish to point out the two marked differences between the S_N2-ra mechanism of the Pd + CH_3Cl system tackled in the present investigation (see also refs 3 and 10) and that of Pd + CH_3F, studied recently.[28] In both cases, there are two competing reaction channels, direct oxidative insertion (OxIn) and an alternative pathway with strong S_N2 character (S_N2-ra). In the first place, however, the C−F bond is much stronger than the C−Cl bond, and activation of the former is associated with significantly higher barriers (via both OxIn and S_N2). Thus, at variance with the situation for Pd + CH_3Cl, the minimum energy path for Pd approaching CH_3F from the backside is, in a sense, redirected from straight nucleophilic substitution and proceeds instead via the relatively low-energy saddle point TS_{CH} for insertion into a C−H bond. Furthermore, for both, Pd + CH_3F and Pd + CH_3Cl, the highest point on the PES of the S_N2-ra pathway has the character of a migrating leaving group, that is, F^- and Cl^-, respectively, that is expelled during the actual substitution process. However, the much higher basicity of F^- compared to Cl^- causes the former, after its expulsion in the actual S_N2 transition state TS_{S_N2} and on its way toward Pd, to abstract a proton, under formation of the intermediate complex IM_{S_N2} between $PdCH_2$ and HF (see ref 28). From the latter, fluoride migrates via transition state TS_{S_N2-ra} toward Pd under formation of the product CH_3PdF. At variance, in the case of Pd + CH_3Cl, the expelled Cl^- leaving group migrates directly to Pd *without abstracting a proton* and, thus, without forming an additional intermediate complex involving the conjugate acid HCl. Thus, the only transition state encountered along the reaction coordinate of the nucleophilic substitution reaction between Pd and CH_3Cl is the one with chloride rearrangement character: TS_{S_N2-ra}.

All species in both reaction pathways have been verified through vibrational analyses to represent equilibrium structures (no imaginary frequencies) or transition states (one imaginary frequency). Furthermore, it has been verified that each transition state connects the stable stationary points as reported.

The geometries obtained with the various LDA and GGA functionals do not show significant mutual discrepancies (see Table S1 and Figure S1 in the Supporting Information). One

Table 2. Relative Energies (in kcal/mol) of the Stationary Points along the Reaction Coordinates of the OxIn- and S_N2-type Pathways for Oxidative Addition of the C−Cl Bond of CH_3Cl to Pd, without (no CPC) and with Counterpoise Correction (with CPC), Computed at Several Levels of ab Initio Theory

		RC_{OxIn}		RC_{S_N2}		TS_{OxIn}		TS_{S_N2-ra}[a]		P	
method	basis set	no CPC	with CPC	no CPC	with CPC	no CPC	with CPC	no CPC	with CPC	no CPC	with CPC
HF	BS1	6.4	6.9	9.9	10.4	28.9	29.4	61.6	62.2	4.3	4.8
	BS2	6.2	6.7	9.8	10.3	28.6	29.2	60.6	61.1	2.0	2.6
	BS2(−)	7.2	7.7	9.8	10.3	29.0	29.6	58.6	59.1	1.6	2.2
	BS2(+)	6.2	6.7			28.4	28.9				
	BS3	6.0	6.5	9.5	10.1	28.2	28.8	59.6	60.2	0.1	0.8
	BS4	6.0	6.5	9.6	10.0	28.2	28.7	59.6	60.1	0.0	0.6
	BS5	5.9	6.4	9.5	10.0	28.1	28.6	59.3	59.9	−0.6	0.0
MP2	BS1	−11.2	−6.4	−5.8	−0.9	3.9	10.4	41.9	47.3	−29.5	−19.7
	BS2	−16.7	−9.8	−10.6	−3.5	−3.1	6.2	37.3	45.1	−39.0	−25.0
	BS2(−)	−10.8	−6.2	−10.1	−3.3	1.7	9.3	36.7	44.4	−34.3	−22.6
	BS2(+)	−17.4	−10.0			−8.5	4.6				
	BS3	−16.9	−13.7	−9.6	−6.5	−2.0	1.8	45.1	48.6	−30.4	−25.0
	BS4	−16.3	−14.2	−7.7	−6.0	−1.1	1.4	46.4	48.6	−29.4	−25.4
	BS5	−16.7	−14.8	−9.0	−7.3	−1.8	0.5	46.7	48.7	−29.1	−25.4
CCSD	BS1	−8.3	−3.5	−4.3	0.7	7.2	13.6	40.0	45.5	−28.7	−19.1
	BS2	−11.3	−5.3	−6.9	−0.7	3.6	11.8	38.2	45.1	−33.8	−21.8
	BS2(−)	−6.5	−2.2	−6.6	−0.5	7.4	14.2	37.4	44.3	−29.9	−19.7
	BS2(+)	−11.9	−5.4			−0.8	10.4				
	BS3	−10.2	−7.3	−5.0	−2.3	6.4	9.7	44.9	48.1	−27.1	−22.4
	BS4	−9.8	−7.8	−4.1	−2.6	7.0	9.3	45.9	48.0	−26.4	−22.8
	BS5	−9.6	−7.9	−4.2	−2.8	6.9	9.1	46.4	48.4	−25.8	−22.5
CCSD(T)	BS1	−11.0	−5.1	−7.0	−0.8	2.1	10.0	35.0	42.0	−33.8	−22.3
	BS2	−14.9	−7.7	−10.3	−2.7	−2.6	7.1	31.7	40.2	−40.7	−26.5
	BS2(−)	−9.4	−4.2	−9.9	−2.5	1.9	10.1	31.1	39.5	−36.1	−24.0
	BS2(+)	−15.6	−7.9			−7.7	5.5				
	BS3	−14.1	−10.3	−8.5	−4.7	0.2	4.7	38.5	43.1	−33.6	−27.6
	BS4	−13.1	−11.0	−6.8	−5.0	1.6	4.2	40.3	43.0	−32.0	−28.0
	BS5	−13.1	−11.2	−7.0	−5.4	1.4	3.8	40.7	43.3	−31.7	−28.0

[a] CCSD(T) procedure not reliable for C−Cl S_N2 transition state, see Section 3.2.

eye-catching, but not essential, difference is the product P computed with VWN and PW91. Here, the methyl group is rotated into an eclipsed instead of a staggered conformation relative to the Pd−Cl bond, at variance with the product geometries for the other functionals. It should be noted, however, that enforcing a staggered geometry will raise the energy by only 0.2 kcal/mol for VWN and a virtually negligible 0.03 kcal/mol for PW91. In fact, the essential physics here is that the methyl group is virtually a free internal rotor.

The C−H bond distance values are very robust with respect to changing the functional, with variations on the order of a few hundredths, or less, of an angstrom. Note that variations in the length of the activated C−Cl bond become larger, up to ca. 0.2 Å in the product, as the reaction progresses. This is in line with the fact that this bond is being broken along the reaction coordinate, which causes the PES to become increasingly soft in this coordinate and, thus, sensitive to changes in the computational method. More pronounced variations are found for the weak Pd−C, Pd−H, and Pd−Cl bonds. This holds especially for the loosely bound reactant complex RC_{S_N2} and the unstable transition state TS_{S_N2-ra}, which for the GGA functionals show fluctuations of up to more than 0.1 Å for Pd−C and Pd−Cl (LDA deviates a bit more, up to 0.5 Å for Pd−Cl). The variations in these bond distances drop to a few hundredths or even a few thousandths of an angstrom as the reaction proceeds to the product in which more stable coordination bonds are formed.

Thus, the various functionals yield essentially the same geometries. Because we found in previous studies on the reaction of Pd with methane and ethane that BLYP performed excellently in terms of the relative energies of stationary points for those model reactions[26,27] and because BLYP is robust and well established, we chose the geometries of this functional, that is, ZORA-BLYP/TZ2P, to compute the ab initio benchmark potential energy surface in the next section.

3.2. Benchmark Energies from ab Initio Calculations. Here, we report the first systematic ab initio calculations into relative energies of the model addition reaction of the C−Cl bond of chloromethane to the Pd atom. This survey is based on geometries of stationary points that were optimized at the ZORA-BLYP/TZ2P level of relativistic DFT (see preceding section and Figure 1). The results of our ab initio computations are collected in Tables 2 and 3 (relative energies and BSSE). Tables S2 and S3 in the Supporting Information show the total energies in atomic units of all species occurring at the stationary points as well as the BSSE for all methods and all stationary points. The reaction profiles

Table 3. Basis Set Superposition Error (BSSE, in kcal/mol) for Pd and CH_3Cl in the Stationary Points along the Reaction Coordinates of the OxIn- and S_N2-type Pathways for Oxidative Addition of the C−Cl Bond of CH_3Cl to Pd, Computed at the CCSD(T) Level of ab Initio Theory

	RC_{OxIn}			RC_{S_N2}			TS_{OxIn}			TS_{S_N2-ra}			P		
basis set	Pd	CH_3Cl	total	Pd	CH_3Cl	total	Pd	CH_3Cl	total	Pd	CH_3Cl	total	Pd	CH_3Cl	total
BS1	5.5	0.4	5.9	5.9	0.3	6.2	7.4	0.5	7.9	6.2	0.7	6.9	10.6	0.9	11.5
BS2	6.8	0.4	7.2	7.2	0.3	7.5	9.2	0.5	9.7	7.8	0.7	8.5	13.2	1.0	14.1
BS2(−)	4.3	0.8	5.1	6.9	0.5	7.4	7.2	1.0	8.2	7.7	0.8	8.5	10.4	1.7	12.1
BS2(+)	7.4	0.3	7.7				12.9	0.3	13.2						
BS3	3.2	0.7	3.8	3.2	0.5	3.7	3.6	0.9	4.5	3.5	1.1	4.6	4.5	1.6	6.1
BS4	1.4	0.8	2.2	1.2	0.6	1.8	1.6	1.0	2.6	1.4	1.3	2.7	2.3	1.7	4.1
BS5	1.1	0.8	1.9	1.1	0.6	1.7	1.3	1.1	2.4	1.2	1.4	2.6	1.8	1.8	3.7

obtained with CCSD(T) are graphically displayed in Figure S2 in the Supporting Information.

We proceed with examining the reaction profiles of the two pathways for the oxidative addition of Pd to the chloromethane C−Cl bond, that is, the energies of the stationary points relative to the reactants Pd and chloromethane (see Table 2 and Figure S2, Supporting Information). At almost all levels of theory except Hartree−Fock, the reaction profiles are characterized by the formation of stable reactant complexes RC_{OxIn} and RC_{S_N2}, where the first one is always lower in energy than the second one, which lead via the transition state for direct oxidative insertion (TS_{OxIn}) or via the transition state for rearrangement after the S_N2 reaction (TS_{S_N2-ra}) to the oxidative addition product (P). Three striking observations can be made: (i) the spread in values of computed relative energies, depending on the level of theory and basis set, is enormous, up to ca. 45 kcal/mol; (ii) the size of the BSSE is also remarkably large, up to ca. 14 kcal/mol; (iii) without counterpoise correction, convergence with basis-set size of the computed energies is still not reached with standard basis sets used routinely in CCSD(T) computations on organometallic and coordination compounds. The lack of any correlation, which is important for this model reaction,[80,81] leads to a complete failure at the HF level, which yields unbound reactant complexes and strongly exaggerated activation barriers: ca. 29 kcal/mol for TS_{OxIn} and ca. 60 kcal/mol for TS_{S_N2-ra}. The activation energies for both pathways drop significantly when electron correlation is introduced. Along HF, CCSD, and CCSD(T) in combination with basis set BS1, for example, the activation barrier for direct oxidative insertion decreases from 28.9 to 7.2 to 2.1 kcal/mol (see Table 2). But also, the correlated CCSD(T) values obtained with basis sets BS1 up to BS3, comparable in quality to standard basis sets such as LANL2DZ[82,83] without or with up to four f functions added, are questionable, if one does not take into account counterpoise correction, as they are obviously not converged as a function of the basis-set size. For example, at CCSD(T)/BS1, the activation energy for direct insertion is 2.1 kcal/mol. This activation energy computed at CCSD(T) drops from 2.1 kcal/mol for basis set BS1 to −2.6 kcal/mol for basis set BS2 in which one f polarization function has been added. Thereafter, along BS2 to BS5, the activation energy increases again, although not monotonically, from −2.6 to 1.4 kcal/mol, as three more sets of *f* functions, an additional set of diffuse p functions, and a set of g functions are added to the basis set of Pd (see Tables 1 and 2).

Next, we note that the BSSE takes on large values in the correlated ab initio methods. At the CCSD(T) level, for example, the BSSE for TS_{OxIn} amounts to 7.9, 9.7, 4.5, 2.6, and 2.4 kcal/mol along the basis sets BS1−BS5 (Table 3), whereas the corresponding BSSE values at HF are only ca. 0.6 kcal/mol. The BSSE increases along the reaction coordinate, that is, going from RC_{OxIn} to TS_{OxIn} to P, or going from RC_{S_N2} to TS_{S_N2-ra} to P. The reason for this is that, along these series of stationary points, the carbon, hydrogen, and chlorine atoms and, thus, their basis functions come closer and begin to surround the palladium atom. This effectively improves the flexibility and polarization of the basis set and, thus, the description of the wave function in the region of the palladium atom. Note that the BSSE stems predominantly from the improvement of the stabilization of palladium as chloromethane ghost functions are added. This contribution to the BSSE quickly reduces as the basis set of palladium is improved, and for the two largest basis sets, BS4 and BS5 (which contain g as well as diffuse p functions on Pd), it is on the same order as the extra stabilization of the chloromethane fragment due to adding palladium ghost functions. Note that the total BSSE at CCSD(T) has been considerably decreased, that is, from 9.7 kcal/mol for BS2 to only 2.4 kcal/mol for BS5 (Table 3), and is, thus, not much larger anymore than the relative energies that we compute, in particular, the OxIn barrier of 1.4 kcal/mol, see CCSD(T)/BS5 in Table 2.

In basis sets BS1, BS2, BS3, BS4, and BS5, mentioned above, we use consistently the same basis sets for all substrate atoms, namely, the uncontracted cc-aug-pVDZ for carbon and hydrogen and cc-aug-pVTZ for chlorine. For the oxidative addition of the methane C−H bond to Pd, it was shown that counterpoise-corrected CCSD(T) relative energies at BS5, that is, using uncontracted cc-aug-pVDZ for C and H, are converged within ca. 1 kcal/mol with respect to extending the basis set for C and H to uncontracted cc-aug-pVTZ.[27] Here, we explore to what extent counterpoise-corrected CCSD(T) relative energies of the Pd + CH_3Cl system are converged if the basis set for C and H is extended from cc-aug-pVDZ in basis set BS2 to cc-aug-pVTZ in the larger basis set BS2(+) (see Table 1). Furthermore, we probe the dependence of the counterpoise-corrected CCSD(T) relative energies on the size of the basis set for Cl by

reducing it from cc-aug-pVTZ in basis set BS2 to cc-aug-pVDZ in basis set BS2(−) (see Table 1). The results for the modified basis sets BS2(−) and BS2(+) are also shown in Tables 2 and 3 below the entry for basis set BS2. It appears that using cc-aug-pVDZ instead of cc-aug-pVTZ for chlorine makes a significant difference for the counterpoise-corrected CCSD(T) relative energies. The barrier for oxidative insertion (TS_{OxIn}), for example, changes from 7.1 to 10.1 kcal/mol, going from BS2 to BS2(−) [see Table 2, CCSD(T) with CPC]. From this, we conclude that using the uncontracted cc-aug-pVTZ basis set for the chlorine in chloromethane is a minimal requirement. The calculations with basis set BS2(+) are extremely expensive and were, therefore, confined to the relative energies of two stationary points: RC_{OxIn} and TS_{OxIn}. In agreement with our earlier finding for Pd + CH_4,[27] extending the basis sets of C and H from cc-aug-pVDZ to cc-aug-pVTZ has little effect on the counterpoise-corrected CCSD(T) relative energies. The barrier for oxidative insertion (TS_{OxIn}), for example, decreases by only 1.5 kcal/mol, from 7.1 to 5.5 kcal/mol, going from BS2 to BS2(+) (see Table 2, CCSD(T) with CPC). We conclude that using uncontracted cc-aug-pVDZ for C and H and uncontracted cc-aug-pVTZ for Cl represents a good compromise between computational efficiency and accuracy in our CCSD(T) computations.

Thus, we have been able to achieve virtual convergence of the CCSD(T) relative energies by using a larger than standard basis set and by correcting for the BSSE through counterpoise correction, see Table 2. The counterpoise-corrected relative energies at CCSD(T) are converged to within some tenths of a kilocalorie per mole. For example, the counterpoise-corrected activation energy for direct oxidative insertion (OxIn) at CCSD(T) amounts to 10.0, 7.1, 4.7, 4.2, and 3.8 kcal/mol.

There are, however, strong indications for one of the species, the transition state of the S_N2 pathway TS_{S_N2-ra}, being problematic in the sense that a single-reference ab initio approach to describing it [e.g., HF, MP2, or CCSD(T)] is not suitable: (i) unlike the situation for the other species, the HOMO and LUMO of TS_{S_N2-ra} are degenerate within a few hundredths of an electronvolt; (ii) in line with this, there is near degeneracy of the singlet and triplet states [$E^{triplet} - E^{singlet} = +1.3, -11.0, +3.7, -5.8$, and -1.6 kcal/mol at BLYP/TZ2P, HF, MP2, CCSD, and CCSD(T); ab initio values obtained with BS5 and CPC], and (iii) importantly, the resulting activation energy of 43.3 kcal/mol at CCSD(T)/BS5 with CPC is also much higher than all barriers obtained with the various density functionals, even those which normally overestimate this type of reaction barrier, such as OLYP. For example, the activation barriers obtained with BLYP, OLYP, and B3LYP are 23.1, 31.9, and 36.3 kcal/mol, respectively, all well below the CCSD(T) value of 43.3 kcal/mol. An analysis of the electronic structure of TS_{S_N2-ra} reveals the physics behind this phenomenon: the species has much of the character of a complex between Cl^- and $PdCH_3^+$. Consequently, the HOMO and LUMO of TS_{S_N2-ra} closely resemble a chlorine 3p atomic orbital (AO), pushed up in energy by the (local) excess of negative charge, and a carbon 2p AO on the methyl fragment in $PdCH_3^+$,

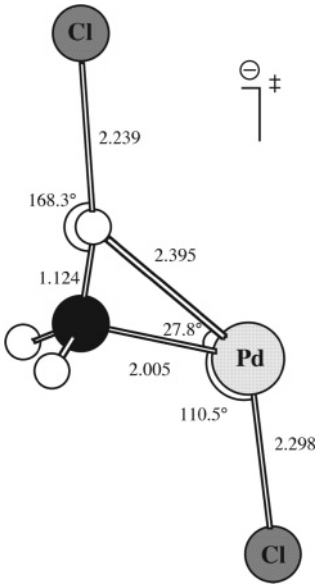


Figure 2. Structure of the S_N2-ra transition state for oxidative addition of the C−Cl bond of CH_3Cl to $PdCl^-$. Geometry optimized at ZORA-BLYP/TZ2P, i.e., with frozen-core approximation.

pulled down in energy by the (local) excess of positive charge: these circumstances clearly promote the occurrence of a single-electron transfer from Cl^- to $PdCH_3^+$. This suggests that the problem may be relieved if the LUMO is destabilized. This can be achieved, for example, by introducing an extra chloride ligand at palladium, which neutralizes the excess positive charge in the $PdCH_3^+$ moiety of TS_{S_N2-ra}. Thus, we have computed and analyzed the corresponding transition state for $PdCl^-$ (instead of Pd) induced C−Cl bond activation, the structure of which is shown in Figure 2. Indeed, all indicators of a pathological situation disappear: (i) there is a clear HOMO−LUMO gap of 0.65 eV at BLYP/TZ2P; (ii) the singlet state is well below the triplet state, and (iii) the counterpoise-corrected CCSD(T)/BS3[84] value for the energy relative to the reactants again agrees perfectly with the BLYP/TZ2P value—both amount to −18.8 kcal/mol (not shown in a Table).[85]

In conclusion, our best estimate, obtained at CCSD(T)/BS5 with CPC, for the kinetic and thermodynamic parameters of the oxidative insertion of Pd into the chloromethane C−Cl bond is −11.2 kcal/mol for the formation of the reactant complex leading to the direct oxidative insertion (OxIn) pathway, −5.4 kcal/mol for the formation of the reactant complex leading to the S_N2 pathway, 3.8 kcal/mol for the activation energy (relative to the reactants) for the OxIn pathway, and −28.0 kcal/mol for the reaction energy (see Table 4). The activation energy of 43.3 kcal/mol for the S_N2 pathway is probably too high for the reasons pointed out above; this value should, therefore, be treated with great precaution and not as a benchmark. If we take into account zero-point vibrational energy (ZPE) effects computed at BLYP/TZ2P, we arrive at −10.8 kcal/mol for the formation of the reactant complex leading to the OxIn pathway, −6.1 kcal/mol for the formation of the reactant complex leading to the S_N2 pathway, 2.7 kcal/mol for the activation energy

Table 4. Relative Energies without (ΔE) and with Zero-Point Vibrational Energy Correction ($\Delta E + \Delta$ZPE) and Relative Enthalpies at 298.15 K (ΔH) of the Stationary Points[a] along the Reaction Coordinates of the OxIn- and S_N2-type Pathways for Oxidative Addition of the C−Cl Bond of CH_3Cl to Pd (in kcal/mol), Computed with Eight Different Density Functionals and the TZ2P Basis Set with Frozen-Core Approximation,[b] and Compared to the ab Initio Benchmark from This Work

	ΔE					$\Delta E + \Delta$ZPE					ΔH				
method	RC_{OxIn}	RC_{S_N2}	TS_{OxIn}	TS_{S_N2-ra}	P	RC_{OxIn}	RC_{S_N2}	TS_{OxIn}	TS_{S_N2-ra}	P	RC_{OxIn}	RC_{S_N2}	TS_{OxIn}	TS_{S_N2-ra}	P
DFT Computations (This Work)[b]															
VWN	−30.1	−25.5	−21.8	15.8	−52.1	−29.6	−26.6	−22.7	11.8	−52.9	−29.9	−27.1	−23.2	11.4	−53.1
BP86	−16.4	−9.3	−5.2	23.4	−36.9	−16.0	−10.2	−6.3	20.4	−37.7	−16.2	−10.6	−6.8	20.1	−37.8
BLYP	−12.9	−5.1	−0.6	23.1	−33.1	−12.5	−5.8	−1.7	20.4	−33.9	−12.7	−6.0	−2.0	20.1	−34.0
PW91	−17.6	−10.8	−6.7	22.7	−37.8	−17.1	−11.6	−7.8	19.8	−38.7	−19.1	−12.0	−10.0	19.5	−39.4
PBE	−17.0	−10.4	−6.1	23.3	−37.1	−16.6	−11.3	−7.2	20.3	−37.9	−16.8	−11.7	−7.7	19.9	−38.7
revPBE	−11.9	−5.0	0.1	26.1	−31.4	−11.6	−5.9	−1.0	23.5	−32.3	−11.8	−6.1	−1.4	23.1	−32.4
RPBE	−11.5	−4.5	0.8	26.1	−30.7	−11.1	−5.3	−0.4	23.3	−31.5	−11.3	−5.6	−0.7	23.0	−31.7
OLYP	−6.8	−0.1	7.0	31.2	−23.5	−6.4	−0.8	5.8	28.7	−24.4	−6.6	−0.9	5.3	28.4	−24.5
Ab Initio Benchmark (This Work)[c]															
CCSD(T)	−11.2	−5.4	3.8	(43.3)[d]	−28.0	−10.8	−6.1	2.7	(40.6)[d]	−28.8					

[a] Geometries and energies computed at the same level of theory. See Figure S1 and Table S1 in the Supporting Information for structures. [b] Relativistic effects treated with ZORA (see Section 2). [c] CCSD(T) benchmark from this work, based on BLYP-optimized geometries. [d] CCSD(T) procedure not reliable for C−Cl S_N2 transition state, see Section 3.2.

(relative to the reactants) for the OxIn pathway, and −28.8 kcal/mol for the reaction energy (see Table 4).

3.3. Validation of DFT. Next, we examine the relative energies of stationary points computed with the LDA functional VWN and the GGA functionals BP86, BLYP, PW91, PBE, revPBE, RPBE, and OLYP in combination with the TZ2P basis set, the frozen-core approximation, and the ZORA to account for relativistic effects. Note that for each density functional we consistently use the geometries optimized with that functional, for example, BP86//BP86 or BLYP//BLYP (see Section 3.1). We focus on the overall activation energy, that is, the difference in energy between the transition state and the separate reactants, which is decisive for the rate of chemical reactions in the gas phase, in particular, if they occur under low-pressure conditions in which the reaction system is (in good approximation) thermally isolated[86,87] (see also Section 2 of ref 88). Relative energies, with and without zero-point vibrational energy correction, as well as relative enthalpies are collected in Table 4 and graphically represented in Figure S3 in the Supporting Information. The performance of the LDA functional VWN and the seven different GGA functionals is assessed by a systematic comparison of the resulting potential energy surfaces with our relativistic four-component CCSD(T) benchmark. It is clear from Table 4 that LDA suffers here from its infamous overbinding, providing barriers that are too low and complexation and reaction energies that are too high. The GGA functionals fall into three groups regarding their agreement with the benchmark results. OLYP clearly underestimates metal−substrate bonding and yields too weakly bound reactant complexes for both pathways, a barrier for the OxIn pathway that is too high by 3.2 kcal/mol, and an insufficiently exothermic reaction energy. The situation is the opposite for BP86, PBE, and PW91, which overestimate metal−substrate bonding, giving rise to too strongly bound reactant complexes, a significantly underestimated barrier for the OxIn pathway (by more than 10 kcal/mol for PW91), and a too exothermic reaction energy. On the other hand, BLYP and the two revisions of PBE, that is, revPBE and RPBE, perform very satisfactorily with reactant complexes in good agreement with the coupled-cluster PES and a relatively small underestimation of the barrier for the OxIn pathway (i.e., by 4.4, 3.7, and 3.0 kcal/mol for BLYP, revPBE, and RPBE, respectively) and somewhat too large reaction energies, but less so than in the case of the group of BP86, PBE, and PW91. Note that all density functionals undershoot to an unusually high extent the CCSD(T) value of the barrier associated with the S_N2 pathway. In Section 3.2, it was pointed out that in this case (i.e., for TS_{S_N2-ra}) the CCSD(T) value tends to be too high and should be treated with great precaution (later on, the issue is again briefly addressed).

We proceed with examining the convergence of the (all-electron) BLYP relative energies of stationary points as the basis set increases along ae-DZ, ae-TZP, ae-TZ2P, and ae-QZ4P, using the ZORA-BLYP/TZ2P geometries, which were also used in the ab initio calculations in the preceding section (see Figure 1). We also investigate the convergence of the BSSE along this series and the effect of using the frozen-core approximation in the calculations discussed in the preceding paragraph. The results are shown in Table 5 and in Figure S4 in the Supporting Information. In the first place, we note that it is valid to use the frozen-core approximation as it has only small effects on the relative energies. This becomes clear if one compares, in Table 5, the frozen-core BLYP/TZ2P results (no CPC: −12.9, −5.1, −0.6, 23.1, and −33.1 kcal/mol for RC_{OxIn}, RC_{S_N2}, TS_{OxIn}, TS_{S_N2-ra}, and P, respectively) with the corresponding all-electron BLYP/ae-TZ2P data (no CPC: −12.8, −5.1, −0.5, 23.1, and −33.5 kcal/mol for RC_{OxIn}, RC_{S_N2}, TS_{OxIn}, TS_{S_N2-ra}, and P, respectively). The frozen-core and all-electron values of the relative energies agree within 0.1−0.4 kcal/mol. Likewise, the BSSE values computed with the frozen-core TZ2P and ae-TZ2P basis sets agree within 0.1 kcal/mol (see Table 5). Next, the issue of basis set convergence is addressed. The data in Table 5 show that the relative energies of stationary points are already converged to within the order of some tenths of a kilocalorie per mole with the ae-TZ2P

Table 5. Relative Energies (in kcal/mol) of the Stationary Points[a] along the Reaction Coordinates of the OxIn- and S_N2-type Pathways for Oxidative Addition of the C−Cl Bond of CH_3Cl to Pd, Computed with BLYP and Four Different Basis Sets with All Electrons Treated Variationally, without (no CPC) and with Counterpoise Correction (with CPC)[b]

	RC_{OxIn}		RC_{S_N2}		TS_{OxIn}		TS_{S_N2-ra}		P	
basis set	no CPC	with CPC	no CPC	with CPC	no CPC	with CPC	no CPC	with CPC	no CPC	with CPC
ae-DZ	−6.5	−4.4	−2.4	0.5	1.9	5.3	16.8	19.0	−32.8	−29.6
ae-TZP	−12.0	−11.9	−4.6	−4.4	0.3	0.5	23.0	23.4	−31.8	−31.4
ae-TZ2P	−12.8	−12.6	−5.1	−4.8	−0.5	−0.2	23.1	23.5	−33.5	−33.0
ae-QZ4P	−13.3	−13.3	−5.3	−5.2	−0.8	−0.7	23.0	23.1	−33.4	−33.3
TZ2P[c]	−12.9	−12.7	−5.1	−4.9	−0.6	−0.4	23.1	23.4	−33.1	−32.7

[a] Geometries optimized at ZORA-BLYP/TZ2P with frozen-core approximation, see Figure 1. [b] Relativistic effects treated with ZORA (see Section 2). [c] Standard frozen-core basis set (see Section 2.1).

basis set. The BSSE drops to 0.5 kcal/mol or less for this basis set and becomes even smaller, that is, 0.1 kcal/mol or less, if one goes to ae-QZ4P (see Table 5: the BSSE is the difference between "no CPC" and "with CPC" values). For example, the activation energy for the OxIn pathway, without counterpoise correction, varies from 1.9 to 0.3 to −0.5 to −0.8 kcal/mol along ae-DZ, ae-TZP, ae-TZ2P, and ae-QZ4P (Table 5, no CPC). The corresponding BSSE amounts to 3.4, 0.2, 0.3, and 0.1 kcal/mol (see Table 5). Note that, in fact, the BSSE is large, that is, a few kilocalories per mole, only for the smallest, ae-DZ, basis set. This is in line with our previous work on the oxidative addition of methane, ethane, and fluoromethane to Pd in which we found that basis-set convergence and elimination of the BSSE are achieved much earlier for DFT (e.g., B3LYP or BLYP) than for correlated ab initio methods [e.g., CCSD(T)].[25−28] In general, correlated ab initio methods depend more strongly on the extent of polarization of the basis set because the polarization functions are essential to generate the configurations through which the wave function can describe the correlation hole. In DFT, on the other hand, the correlation hole is built into the potential and the energy functional and polarization functions mainly play the much less delicate role of describing polarization of the electron density. In conclusion, the TZ2P basis in combination with the frozen-core approximation yields an efficient and accurate (i.e., within 1 kcal/mol) description of the relative energies of our stationary points.

Finally, on the basis of the ZORA-BLYP/TZ2P geometries discussed above, we have computed the relative energies of stationary points along the PES for various LDA, GGA, meta-GGA, and hybrid functionals in combination with the all-electron ae-TZ2P basis set and ZORA for relativistic effects. This was done in a post-SCF manner, that is, using density functionals with the electron density obtained at ZORA-BLYP/ae-TZ2P. The performance of the density functionals is discussed by comparing the resulting potential energy surfaces with that of the ab initio [CCSD(T)] benchmark discussed above. The results of this survey are collected in Table 6, which shows energies relative to the separate reactants (R).

For clarity, we wish to point out that the above procedure for computing the relative energies shown in Table 6 differs in three respects from that used for computing the relative energies with the LDA functional and the seven GGA functionals shown in Table 4: (i) an all-electron approach is used instead of the frozen-core approximation; (ii) for all density functionals, the BLYP optimized geometries are used instead of geometries optimized with the same functional, and (iii) for all functionals, the BLYP electron density is used for computing the energy instead of the electron density corresponding to that functional. The effect of going from frozen-core (TZ2P) to all-electron calculations (ae-TZ2P), that is, point i, is small, causing a stabilization of 0.3 kcal/mol or less, and has already been discussed above. The differences between the values in Tables 4 and 6 derive mainly from the combined effect of points ii and iii, which causes, considering the GGA functionals, a destabilization of up to 1.0 kcal/mol (for the PBE and OLYP transition state for the OxIn pathway) of the relative energies if one goes from Table 4 to Table 6. Both effects are on the order of a few tenths of a kilocalorie per mole up to maximally 1 kcal/mol and, for the different GGA functionals and stationary points, contribute to this destabilization with varying relative importance. For example, for TS_{OxIn}, the single-point approach contributes generally somewhat more (0.6−1.0 kcal/mol) to this destabilization than the post-SCF approach (up to 0.3 kcal/mol). This has been assessed by computing the relative energies of stationary points using approximation ii but not iii, that is, computing them with the electron density corresponding to the density functional under consideration but with the BLYP geometries; the resulting values are provided in parentheses in Table 6. In conclusion, for the GGA functionals, the combined effect of approximations i−iii on the relative energies of stationary points is on the order of a few tenths of a kilocalorie per mole with an upper limit of 1 kcal/mol.

Now, we extend our survey to the full range of energy density functionals that, except for LDA and the seven GGAs discussed above, have been implemented in the ADF program in a post-SCF manner. For all 26 density functionals, we have computed the mean absolute error in the relative energies of the reactant complexes, transition states, and product and the error in the barriers, that is, the relative energy of the transition states, as compared with the CCSD(T) benchmark (see Table 6). In Section 3.2, we have pointed out that the CCSD(T) relative energy for the S_N2 transition state TS_{S_N2-ra} is unreliable and must be treated with great precaution. Indeed, for this particular species, the counterpoise-corrected CCSD(T)/BS5 value of the relative energy exceeds the corresponding values obtained with the various density functionals to an unusually great extent, even those which normally overestimate this type of reaction

Table 6. Energies (in kcal/mol) Relative to the Separate Reactants (R) of the Stationary Points[a] along the Reaction Coordinates of the OxIn- and S_N2-type Pathways for Oxidative Addition of the C−Cl Bond of CH_3Cl to Pd and Dissociation Energy of CH_3Cl into a Methyl Radical and Chlorine Atom (D_{CCl}), Computed for 26 Different Density Functionals with the ae-TZ2P Basis Set with All Electrons Treated Variationally,[b] and Compared to the ab Initio Benchmark from This Work.

method	RC_{OxIn}	RC_{S_N2}	TS_{OxIn}	TS_{S_N2-ra}	P	mean abs. err.[c]	mean abs. err., excl. TS_{S_N2-ra}[d]	err. in OxIn barr.[c]	err. in S_N2-ra barr.[c]	D_{CCl}	err. in D_{CCl}[e]
				LDA							
VWN	−27.6 (−27.7)	−20.7 (−21.0)	−18.2 (−18.4)	18.7 (7.7)	−50.2 (−50.3)	20.1	19.0	−22.0	−24.6	106.2	25.0
				GGA							
BP86	−16.0 (−16.0)	−8.6 (−8.7)	−4.5 (−4.5)	23.8 (23.8)	−37.0 (−37.0)	9.0	6.3	−8.3	−19.5	86.0	4.9
BLYP	−12.8	−5.1	−0.5	23.1	−33.5	6.4	2.9	−4.2	−20.2	82.1	1.0
B88xBR89c	−13.9	−5.8	−0.4	22.6	−35.9	7.2	3.8	−4.2	−20.7	83.4	2.2
PW91	−17.2 (−17.2)	−10.1 (−10.1)	−5.8 (−5.8)	23.0 (23.1)	−38.0 (−37.9)	10.1	7.6	−9.6	−20.3	88.9	7.7
PBE	−16.4 (−16.6)	−9.5 (−9.6)	−5.1 (−5.1)	23.6 (23.7)	−37.2 (−37.1)	9.4	6.9	−8.9	−19.7	88.9	7.7
FT97	−12.7	−10.5	3.8	23.5	−36.9	7.1	3.9	0.0	−19.8	84.7	3.5
revPBE	−11.6 (−11.8)	−4.6 (−4.8)	0.7 (0.7)	26.3 (26.2)	−31.7 (−31.7)	5.0	2.0	−3.1	−17.0	83.5	2.4
HCTH/93	−6.6	0.4	8.0	32.1	−22.7	6.2	4.9	4.2	−11.2	83.0	1.9
RPBE	−11.1 (−11.3)	−4.1 (−4.3)	1.4 (1.3)	26.3 (26.2)	−30.9 (−30.9)	4.7	1.7	−2.4	−17.0	82.9	1.8
BOP	−9.6	−1.9	3.4	25.4	−29.9	5.0	1.8	−0.3	−17.9	81.8	0.6
HCTH/120	−11.0	−3.9	3.0	28.5	−27.5	3.6	0.8	−0.8	−14.8	84.3	3.2
HCTH/147	−10.4	−3.3	3.6	29.1	−27.0	3.7	1.0	−0.2	−14.2	84.3	3.1
HCTH/407	−7.8	−1.3	8.0	30.8	−22.7	5.9	4.2	4.2	−12.4	83.2	2.1
OLYP	−6.0 (−6.5)	0.7 (0.2)	8.0 (7.7)	31.9 (31.3)	−23.3 (−23.7)	6.3	5.0	4.2	−11.4	83.6	2.4
				Meta-GGA							
BLAP3	−7.7	0.2	7.5	27.7	−26.7	5.9	3.5	3.7	−15.6	85.3	4.1
VS98	−14.2	−8.9	−5.7	25.7	−33.5	7.8	5.4	−9.5	−17.6	81.2	0.1
KCIS	−13.7	−6.7	−1.6	26.0	−34.7	6.6	4.0	−5.4	−17.3	85.4	4.2
PKZB	−12.5	−5.2	−0.5	26.6	−34.4	5.8	3.0	−4.3	−16.7	83.6	2.5
Bmτ1	−7.4	0.5	7.9	27.3	−26.4	6.2	3.8	4.1	−16.0	83.7	2.5
OLAP3	−0.9	6.1	16.0	36.5	−16.5	10.4	11.3	12.2	−6.8	86.8	5.6
TPSS	−14.4	−6.7	−3.7	25.0	−36.9	7.9	5.3	−7.5	−18.3	84.2	3.0
				Hybrid							
B3LYP	−9.3	−3.1	5.4	36.3	−26.5	2.8	1.8	1.6	−7.0	81.2	0.1
O3LYP	−5.4	0.5	10.1	40.4	−19.6	5.9	6.6	6.3	−2.9	85.7	4.5
X3LYP	−9.7	−3.7	5.0	36.9	−26.7	2.4	1.4	1.2	−6.3	81.7	0.6
TPSSh	−12.3	−5.4	−0.3	31.6	−32.9	4.4	2.6	−4.1	−11.7	83.2	2.0
				Ab Initio Benchmark (This Work)[f]							
CCSD(T)	−11.2	−5.4	3.8	(43.3)[g]	−28.0					81.2	

[a] Geometries optimized at ZORA-BLYP/TZ2P with frozen-core approximation, see Figure 1. [b] Computed post-SCF using the BLYP electron density, unless stated otherwise. Values in parentheses computed self-consistently, i.e., with the potential and electron-density corresponding to the energy functional indicated. Relativistic effects treated with ZORA (see Section 2). [c] Mean absolute error for the energies of the five stationary points RC_{OxIn}, RC_{S_N2}, TS_{OxIn}, TS_{S_N2-ra}, and P relative to the separate reactants (R) and error in the overall barriers, i.e., in the energy of TS_{OxIn} and TS_{S_N2-ra}, respectively, relative to R, compared with the CCSD(T) benchmark from this work. [d] Mean absolute error corresponding to footnote c but excluding stationary point TS_{S_N2-ra}. [e] Error in the dissociation energy of the C−Cl bond in chloromethane, compared with the CCSD(T) benchmark from this work. [f] CCSD(T) benchmark from this work, based on BLYP-optimized geometries. [g] CCSD(T) procedure not reliable for C−Cl S_N2 transition state, see Section 3.2.

barrier, such as OLYP (see Table 6). For comparison, deviations of the DFT barriers are significantly smaller for the transition state of the OxIn pathway, and they also show the well-known scattering of individual values somewhat above and below the CCSD(T) benchmark value. Thus, for all density functionals except O3LYP, the mean absolute error between the DFT and CCSD(T) relative energies of the stationary points along the OxIn and S_N2 PESs drops significantly if one excludes the S_N2-ra transition state; Table 6 displays both values in the columns "mean abs. err." and "mean abs. err. excl. TS_{S_N2-ra}". In the following, we discuss the latter. Both the mean absolute error and the error in the OxIn barrier drop significantly if one goes from LDA (mean absolute error = 19.0 kcal/mol), which, as mentioned above, suffers from its infamous overbinding, to GGA functionals (mean absolute error = 0.8−7.6 kcal/mol). However, no significant improvement occurs if one goes from GGA to the more recently developed meta-GGA functionals (mean absolute error = 3.0−11.3 kcal/mol) and hybrid functionals (mean absolute error = 1.4−6.6 kcal/mol). The best overall agreement with the ab initio benchmark PES is achieved by functionals of the GGA (HCTH/120), meta-GGA (PKZB), and hybrid DFT type (X3LYP), with mean absolute errors of 0.8−3.0 kcal/mol and errors in the OxIn barrier ranging from −4.3 to 1.2 kcal/mol. Interestingly, the well-known BLYP functional compares very reasonably with an only slightly larger mean absolute error of 2.9 kcal/mol and an underestimation of the OxIn barrier of −4.2 kcal/mol. The OLYP functional overestimates the OxIn barrier by the same amount, 4.2 kcal/mol, but has a larger mean absolute error

Table 7. Relative Energies (in kcal/mol) of the Stationary Points along the Reaction Coordinate for the Oxidative Addition of Pd to the C−H, C−C, C−F, and C−Cl Bonds, Computed with CCSD(T), BLYP, OLYP, and B3LYP[a]

		reactant complex				transition state				product			
activated bond		CCSD(T)	BLYP	OLYP	B3LYP	CCSD(T)	BLYP	OLYP	B3LYP	CCSD(T)	BLYP	OLYP	B3LYP
C−H	OxIn	−8.1	−6.7	−0.7	−4.9	5.8	3.9	10.3	10.3	0.8	−3.6	5.3	4.6
C−C	OxIn	−10.8	−6.7	−0.5	−4.9	19.4	18.5	26.8	25.2	−4.5	−9.5	1.6	0.2
C−F	OxIn	−5.3	−5.4	0.3	−3.4	27.8	17.7	26.0	26.9	−6.4	−16.3	−6.2	−7.0
	S_N2					37.5	30.1	38.5	40.2				
C−Cl	OxIn	−11.2	−12.8	−6.0	−9.3	3.8	−0.5	8.0	5.4	−28.0	−33.5	−23.3	−26.5
	S_N2	−5.4	−5.1	0.7	−3.1	(43.3)[b]	23.1	31.9	36.3				

[a] Geometries optimized at ZORA-BLYP/TZ2P with frozen-core approximation. BLYP, OLYP, and B3LYP results calculated with the ae-TZ2P basis set with all electrons treated variationally and post-SCF using the BLYP electron density. CCSD(T) results calculated with relativistic four-component method. For details, see refs 25, 26 (C−H), 27 (C−C), 28 (C−F), and this work (C−Cl). [b] CCSD(T) procedure not reliable for C−Cl S_N2 transition state, see Section 3.2.

of 5.0 kcal/mol (in the case of C−F bond activation,[28] OLYP performs better than BLYP). The hybrid functionals B3LYP and X3LYP perform remarkably well, with overestimations of the barrier of only 1.6 and 1.2 kcal/mol and mean absolute errors of only 1.8 and 1.4 kcal/mol, respectively.

We have verified to what extent errors made, for example, by BLYP or B3LYP, originate from a failure in describing the C−Cl bond dissociation. To this end, we have first computed an ab initio benchmark for the C−Cl bond strength, that is, the dissociation energy D_{CCl}, associated with the reaction $H_3C-Cl \rightarrow CH_3^{\bullet} + Cl^{\bullet}$ at the same levels of theory as we did for the PES of the oxidative addition of the chloromethane C−Cl bond to Pd. This was done again using the BLYP-optimized geometries, which yield a C−H bond length of 1.084 Å for the D_{3h} symmetric methyl radical. Thus, we arrive at a dissociation energy of 81.2 kcal/mol at CCSD(T) with basis set BS5 and with counterpoise correction (HF, 60.0; MP2, 86.0; and CCSD, 78.2 kcal/mol; for details, see Table S5 in the Supporting Information), in nice agreement with the experimental value for the enthalpy at 0 K, namely, 82.04 ± 0.26 kcal/mol.[89] Most functionals are able to describe the dissociation energy reasonably well, yielding errors, compared with the CCSD(T) benchmark, on the order of a few kilocalories per mole. For BLYP and B3LYP, the dissociation energy D_{CCl} is overestimated by only 1.0 and 0.1 kcal/mol, respectively (see Table 6). In conclusion, the underestimation of the activation energy by BLYP cannot be ascribed to a failure in describing C−Cl bond dissociation (in fact, the slight error in the latter works in the opposite direction and should raise the value of the barrier). Rather, it may be related to the overbinding between Pd and the methyl and chloride ligands (compare relative energies for P in Table 6).

3.4. Comparison of C−H, C−C, C−F, and C−Cl Bond Activation. Finally, we have carried out a comprehensive comparison of the ab initio CCSD(T) benchmark PESs as well as the corresponding BLYP, OLYP, and B3LYP density functional results for the palladium-induced activation of methane C−H (OxIn),[25,26] ethane C−C (OxIn),[27] fluoromethane C−F (OxIn and S_N2),[28] and the C−Cl bond (this work, OxIn and S_N2) using the same computational details throughout. The energies of all stationary points relative to the reactants are collected in Table 7. Trends in activation energies are graphically displayed in Figure 3 in which the questionable CCSD(T) value for the S_N2 transition state for C−Cl activation has been left out (see also Section 3.2).

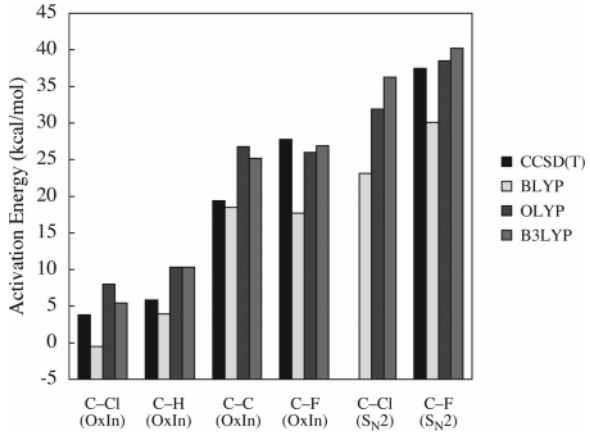


Figure 3. Activation energies (in kcal/mol) for the oxidative addition of Pd to various C−X bonds, computed with CCSD(T), BLYP, OLYP, and B3LYP. For computational details, see the footnotes of Table 7.

It is clear, especially from Figure 3, that all important features of the CCSD(T) benchmark potential energy surfaces for palladium-induced C−H, C−C, C−F, and C−Cl bond activation are reproduced by important functionals such as BLYP, OLYP, and B3LYP. On the other hand, a more detailed look also shows that none of these functionals is the "best one" for each individual model reaction. For example, BLYP performs best in the case of C−H and C−C bond activation whereas OLYP and B3LYP overestimate the barrier (compare values in Table 7). But, in the case of C−F bond activation, the BLYP functional underestimates the barriers of both OxIn and S_N2 pathways while OLYP and B3LYP perform very satisfactorily [Table 7, OxIn pathway: CCSD(T), 27.8 kcal/mol; BLYP, 17.7 kcal/mol; OLYP, 26.0 kcal/mol; B3LYP, 26.9 kcal/mol; S_N2-pathway: CCSD(T), 37.5 kcal/mol; BLYP, 30.1 kcal/mol; OLYP, 38.5 kcal/mol; B3LYP, 40.2 kcal/mol]. For the C−Cl bond, as described above, the OxIn barrier is only slightly underestimated by BLYP and overestimated by OLYP and B3LYP. Nevertheless, they all agree with the CCSD(T) benchmark that, for example, the activation energies for oxidative addition increase in the order C−Cl (OxIn) < C−H (OxIn) < C−C (OxIn) ≲ C−F (OxIn) < C−Cl (S_N2-ra; no reliable benchmark) < C−F (via S_N2-ra), see Figure 3.

4. Conclusions

We have computed an ab initio benchmark for the archetypal oxidative addition of the chloromethane C−Cl bond to palladium that derives from a hierarchical series of relativistic methods and highly polarized basis sets for the palladium atom, up to the counterpoise corrected, four-component spin-free Dirac-Coulomb CCSD(T)/(24s16p13d+4f+p+g) level, which is converged with respect to the basis-set size within 1 kcal/mol. Our findings stress the importance of sufficient higher-angular momentum polarization functions, f and g, as well as counterpoise correction for obtaining reliable activation energies.

This benchmark is used to evaluate the performance of 26 relativistic (ZORA) density functionals for describing relative energies of stationary points on the potential energy surface. Excellent agreement with our ab initio benchmark for energies relative to the reactants is achieved by functionals of the GGA, meta-GGA, and hybrid DFT approaches, each of which have a representative in the top three, with mean absolute errors as small as 3.0 kcal/mol or less. All theoretical methods used reveal the existence of two possible reaction mechanisms for oxidative addition: direct oxidative insertion (OxIn) with a barrier that is at least some 20 kcal/mol lower than that of an alternative S_N2 pathway. Interestingly, the well-known BLYP functional still performs satisfactorily with a mean absolute error of 2.9 kcal/mol and an underestimation of the OxIn barrier by −4.2 kcal/mol. Note that the much advocated B3LYP hybrid functional also performs remarkably well, with a mean absolute error of 1.8 kcal/mol and an overestimation of the OxIn barrier by only 1.6 kcal/mol.

Finally, a comprehensive comparison of the present (C−Cl) and previous studies[25−28] shows that all important features of the CCSD(T) benchmark potential energy surfaces for palladium-induced C−H, C−C, C−F, and C−Cl bond activation are reproduced by important functionals such as BLYP, OLYP, and B3LYP. Thus, while none of these functionals is the "best one" for each individual model reaction, they all agree with the CCSD(T) benchmark that, for example, the activation energies for oxidative addition increase in the order C−Cl (OxIn) < C−H (OxIn) < C−C (OxIn) $\lesssim$ C−F (OxIn) < C−Cl (S_N2-ra; no reliable benchmark) < C−F (via S_N2-ra). Our DFT results have been verified to be converged with the basis-set size at ZORA-BLYP/TZ2P and to be unaffected by the frozen-core approximation for the core shells of carbon (1s), chlorine (1s2s2p), and palladium (1s2s2p3s3p3d). We consider this a sound and efficient approach for the routine investigation of catalytic bond activation, also in larger, more realistic model systems.

Acknowledgment. We thank the Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO-CW and NWO-NCF) for financial support. We thank Ivan Infante for helpful discussions.

Supporting Information Available: Structures of stationary points optimized with LDA and various GGA functionals, total energy and BSSE of all stationary points involved at all levels of ab initio theory applied in the computations, and reaction profiles obtained with CCSD(T) and various GGA functionals. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Frisch, A. C.; Beller, M. *Angew. Chem.* **2005**, *117*, 680.

(2) Senn, H. M.; Ziegler, T. *Organometallics* **2004**, *23*, 2980.

(3) Bickelhaupt, F. M.; Ziegler, T.; von Ragué Schleyer, P. *Organometallics* **1995**, *14*, 2288.

(4) Albert, K.; Gisdakis, P.; Rösch, N. *Organometallics* **1998**, *17*, 1608.

(5) Sundermann, A.; Uzan, O.; Martin, J. M. L. *Chem.—Eur. J.* **2001**, *7*, 1703.

(6) Diefenbach, A.; Bickelhaupt, F. M. *J. Chem. Phys.* **2001**, *115*, 4030.

(7) Smurnyi, E. D.; Gloriozov, I. P.; Ustynuk, Y. A. *Russ. J. Phys. Chem.* **2003**, *77*, 1699.

(8) Reinhold, M.; McGrady, J. E.; Perutz, R. N. *J. Am. Chem. Soc.* **2004**, *126*, 5268.

(9) Diefenbach, A.; Bickelhaupt, F. M. *J. Phys. Chem. A* **2004**, *108*, 8460.

(10) Diefenbach, A.; de Jong, G. Th.; Bickelhaupt, F. M. *J. Chem. Theory Comput.* **2005**, *1*, 286.

(11) Diefenbach, A.; de Jong, G. Th.; Bickelhaupt, F. M. *Mol. Phys.* **2005**, *103*, 995.

(12) Dedieu, A. *Chem. Rev.* **2000**, *100*, 543.

(13) Kiplinger, J. L.; Richmond, T. G.; Osterberg, C. E. *Chem. Rev.* **1994**, *94*, 373.

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

(15) Cizek, J. *J. Chem. Phys.* **1966**, *45*, 4256.

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

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

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

(19) Baker, J.; Muir, M.; Andzelm, J. *J. Chem. Phys.* **1995**, *102*, 2063.

(20) Barone, V.; Adamo, C. *J. Chem. Phys.* **1996**, *105*, 11007.

(21) Thümmel, H. T.; Bauschlicher, C. W., Jr. *J. Phys. Chem. A* **1997**, *101*, 1188.

(22) Bach, R. D.; Glukhovtsev, M. N.; Gonzales, C. *J. Am. Chem. Soc.* **1998**, *120*, 9902.

(23) Gritsenko, O. V.; Ensing, B.; Schippers, P. R. T.; Baerends, E. J. *J. Phys. Chem. A* **2000**, *104*, 8558.

(24) Poater, J.; Solà, M.; Duran, M.; Robles, J. *Phys. Chem. Chem. Phys.* **2002**, *4*, 722.

(25) de Jong, G. Th.; Solà, M.; Visscher, L.; Bickelhaupt, F. M. *J. Chem. Phys.* **2004**, *121*, 9982.

(26) de Jong, G. Th.; Geerke, D. P.; Diefenbach, A.; Bickelhaupt, F. M. *Chem. Phys.* **2005**, *313*, 261.

(27) de Jong, G. Th.; Geerke, D. P.; Diefenbach, A.; Solà, M.; Bickelhaupt, F. M. *J. Comput. Chem.* **2005**, *26*, 1006.

(28) de Jong, G. Th.; Bickelhaupt, F. M. *J. Phys. Chem. A* **2005**, *109*, 9685.

(29) Hohenberg, P.; Kohn, W. *Phys. Rev.* **1964**, *136*, B864.

(30) Kohn, W.; Sham, L. J. *Phys. Rev.* **1965**, *140*, A1133.

(31) Parr, R. G.; Yang, W. *Density-Functional Theory of Atoms and Molecules*; Oxford University Press: New York, 1989.

(32) Baerends, E. J.; Ellis, D. E.; Ros, P. *Chem. Phys.* **1973**, *2*, 41.

(33) Fonseca Guerra, C.; Snijders, J. G.; te Velde, G.; Baerends, E. J. *Theor. Chem. Acc.* **1998**, *99*, 391.

(34) te Velde, G.; Bickelhaupt, F. M.; Baerends, E. J.; Fonseca Guerra, C.; van Gisbergen, S. J. A.; Snijders, J. G.; Ziegler, T. *J. Comput. Chem.* **2001**, *22*, 931.

(35) Baerends, E. J.; Autschbach, J. A.; Bérces, A.; Bo, C.; Boerrigter, P. M.; Cavallo, L.; Chong, D. P.; Deng, L.; Dickson, R. M.; Ellis, D. E.; Fan, L.; Fischer, T. H.; Fonseca Guerra, C.; van Gisbergen, S. J. A.; Groeneveld, J. A.; Gritsenko, O. V.; Grüning, M.; Harris, F. E.; van den Hoek, P.; Jacobsen, H.; van Kessel, G.; Kootstra, F.; van Lenthe, E.; Osinga, V. P.; Patchkovskii, S.; Philipsen, P. H. T.; Post, D.; Pye, C. C.; Ravenek, W.; Ros, P.; Schipper, P. R. T.; Schreckenbach, G.; Snijders, J. G.; Sola, M.; Swart, M.; Swerhone, D.; te Velde, G.; Vernooijs, P.; Versluis, L.; Visser, O.; van Wezenbeek, E.; Wiesenekker, G.; Wolff, S. K.; Woo, T. K.; Ziegler, T. *ADF2002.03*; SCM, Theoretical Chemistry, Vrije Universiteit: Amsterdam, The Netherlands, 2002.

(36) Vosko, S. H.; Wilk, L.; Nusair, M. *Can. J. Phys.* **1980**, *58*, 1200.

(37) Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098.

(38) Perdew, J. P. *Phys. Rev. B* **1986**, *33*, 8822.

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

(40) Perdew, J. P. In *Electronic Structure of Solids '91*; Ziesche, P., Eschrig, H., Eds.; Akademie Verlag: Berlin, 1991.

(41) Perdew, J. P.; Wang, Y. *Phys. Rev. B* **1992**, *45*, 13244.

(42) 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.

(43) Perdew, J. P.; Chevary, J. A.; Vosko, S. H.; Jackson, K. A.; Pederson, M. R.; Singh, D. J.; Fiolhais, C. *Phys. Rev. B* **1993**, *49*, 4978(E).

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

(45) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1997**, *78*, 1396(E).

(46) Zhang, Y.; Yang, W. *Phys. Rev. Lett.* **1998**, *80*, 890.

(47) Hammer, B.; Hansen, L. B.; Nørkov, J. K. *Phys. Rev. B* **1999**, *59*, 7413.

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

(49) van Lenthe, E.; Baerends, E. J.; Snijders, J. G. *J. Chem. Phys.* **1994**, *101*, 9783.

(50) Visscher, L.; Lee, T. J.; Dyall, K. G. *J. Chem. Phys.* **1996**, *105*, 8769.

(51) Jensen, H. J. A.; Saue, T.; Visscher, L. *DIRAC*, release 4.0; Syddansk Universitet: Odense, Denmark, 2004.

(52) Dyall, K. G. *J. Chem. Phys.* **1994**, *100*, 2118.

(53) Visscher, L. *Theor. Chem. Acc.* **1997**, *98*, 68.

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

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

(56) Visscher, L.; Aerts, P. J. C.; Visser, O.; Nieuwpoort, W. C. *Int. J. Quantum Chem.* **1991**, *25*, 131.

(57) Ehlers, A. W.; Böhme, M.; Dapprich, S.; Gobbi, A.; Höllwarth, A.; Jonas, V.; Köhler, K. F.; Stegmann, R.; Veldkamp, A.; Frenking, G. *Chem. Phys. Lett.* **1993**, *208*, 111.

(58) Langhoff, S. R.; Petterson, L. G. M.; Bauschlicher, C. W., Jr. *J. Chem. Phys.* **1987**, *86*, 268.

(59) Osanai, Y.; Sekiya, M.; Noro, T.; Koga, T. *Mol. Phys.* **2003**, *101*, 65.

(60) Becke, A. D. *J. Chem. Phys.* **1988**, *88*, 1053.

(61) Becke, A. D.; Roussel, M. R. *Phys. Rev. A* **1989**, *39*, 3761.

(62) Filatov, M.; Thiel, W. *Mol. Phys.* **1997**, *91*, 847.

(63) Hamprecht, F. A.; Cohen, A. J.; Tozer, D. J.; Handy, N. C. *J. Chem. Phys.* **1998**, *109*, 6264.

(64) Tsuneda, T.; Suzumura, T.; Hirao, K. *J. Chem. Phys.* **1999**, *110*, 10664.

(65) Boese, A. D.; Doltsinis, N. L.; Handy, N. C.; Sprik, M. *J. Chem. Phys.* **2000**, *112*, 1670.

(66) Boese, A. D.; Handy, N. C. *J. Chem. Phys.* **2001**, *114*, 5497.

(67) Proynov, E. I.; Sirois, S.; Salahub, D. R. *Int. J. Quantum Chem.* **1997**, *64*, 427.

(68) van Voorhis, T.; Scuseria, G. E. *J. Chem. Phys.* **1998**, *109*, 400.

(69) Krieger, J. B.; Chen, J.; Iafrate, G. J.; Savin, A. In *Electron Correlation and Material Properties*; Gonis, A., Kioussis, N., Eds.; Plenum: New York, 1999.

(70) Perdew, J. P.; Kurth, S.; Zupan, A.; Blaha, P. *Phys. Rev. Lett.* **1999**, *82*, 2544.

(71) Perdew, J. P.; Kurth, S.; Zupan, A.; Blaha, P. *Phys. Rev. Lett.* **1999**, *82*, 5179(E).

(72) Proynov, E.; Chermette, H.; Salahub, D. R. *J. Chem. Phys.* **2000**, *113*, 10013.

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

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

(75) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648.

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

(77) Cohen, A. J.; Handy, N. C. *Mol. Phys.* **2001**, *99*, 607.

(78) Xu, X.; Goddard, W. A., III. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 2673.

(79) Hertwig, R. H.; Koch, W. *Chem. Phys. Lett.* **1997**, *268*, 345.

(80) Frenking, G.; Antes, I.; Böhme, M.; Dapprich, S.; Ehlers, A. W.; Jonas, V.; Neuhaus, A.; Otto, M.; Stegmann, R.; Veldkamp, A.; Vyboishchikov, S. F. Pseudopotential calculations of transition metal compounds. In *Reviews in*

Computational Chemistry; Lipkowitz, K. B., Boyd, D. B., Eds.; VCH Publishers Inc.: New York, 1996; Vol. 8, p 63.

(81) Cundari, T. R.; Benson, M. T.; Lutz, M. L. Effective core potential approaches to the chemistry of heavier elements. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; VCH Publishers Inc.: New York, 1996; Vol. 8, p 145.

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

(83) Jonas, V.; Frenking, G.; Reetz, M. T. *J. Comput. Chem.* **1992**, *13*, 919.

(84) Our computational resources do not allow for larger basis sets than BS3 in the case of the $PdCl^- + CH_3Cl$ model reaction system. This basis set should, however, yield relative energies that are reasonably converged with basis-set size, as can be seen for other stationary points in Table 2.

(85) Counterpoise-corrected relative energies of transition state TS_{S_N2-ra} of $PdCl^- + CH_3Cl$ are −10.2, −13.8, −17.7, and −18.8 kcal/mol at HF, MP2, CCSD, and CCSD(T), respectively, in combination with basis set BS3. For details, see Table S4 in the Supporting Information.

(86) Nibbering, N. M. M. *Adv. Phys. Org. Chem.* **1988**, *24*, 1.

(87) Carroll, J. J.; Weisshaar, J. C. *J. Am. Chem. Soc.* **1993**, *115*, 800.

(88) Bickelhaupt, F. M. *Mass Spectrom. Rev.* **2001**, *20*, 347.

(89) Johnson, R. D., III. NIST Computational Chemistry Comparison and Benchmark Database, NIST Standard Reference Database Number 101, Release 11. http://srdata.nist.gov/cccbdb (accessed June 2005).

CT050254G

JCTC Journal of Chemical Theory and Computation

Theoretical Study of the Structure and Properties of $[(\eta^5\text{-}C_5Me_4H)_2Zr]_2(\mu^2,\eta^2,\eta^2\text{-}N_2)$

Petia Bobadova-Parvanova, David Quinonero-Santiago,† Keiji Morokuma,* and Djamaladdin G. Musaev*

Cherry L. Emerson Center for Scientific Computation and Department of Chemistry, Emory University, Atlanta, Georgia 30322

Received October 18, 2005

Abstract: Recently Pool et al. [Pool, J. A.; Lobkovsky, E.; Chirik, P. J. *Nature* **2004**, *427*, 527.] showed that the $[(\eta^5\text{-Cp}')_2Zr]_2(\mu^2,\eta^2,\eta^2\text{-}N_2)$, Cp′ = η^5-C_5Me_4H, complex is promising for dinitrogen hydrogenation. In the present study we examine computationally the structure and relative energies of different possible positional isomers of this dimer complex as well as different isomers of the monomer complex $(Cp'_2Zr)(\eta^2\text{-}N_2)$. The relative stability of isomers of the monomer is determined by the electrostatic repulsion between the negatively charged N atoms of the N_2 molecule and the negatively charged C atoms of the Cp′ ring that are bound to H. Substitution of H atoms by methyl groups significantly changes the charge distribution in Cp rings, increases the negative charge of C_H atom, and affects the relative stability of the isomers. On the other hand, competition between the electrostatic effects and the steric repulsion determines the relative energy of the positional isomers of the dimer $(Cp'_2Zr)_2(\mu^2,\eta^2,\eta^2\text{-}N_2)$.

Introduction

The activation of molecular nitrogen to produce nitrogen-containing compounds under mild conditions is still one of the most challenging tasks of chemistry.[1] Due to its nonpolar and strong triple bond, molecular nitrogen has very low reactivity. Extreme conditions are needed to convert N_2 into practically useful nitrogen-containing compounds. The Haber-Bosch process, which has been used for almost a century to produce ammonia from dihydrogen and dinitrogen molecules, requires temperatures ranging from 400 to 650 °C and pressure ranging from 200 to 400 atm. [2] Numerous efforts have been made to develop new catalysts, which would facilitate activation of dinitrogen under milder conditions. One of the most promising of these was the synthesis of the dinuclear Zr complex $\{[P_2N_2]Zr\}_2(\mu_2,\eta^2,\eta^2\text{-}N_2)$, where P_2N_2=PhP$(CH_2SiMe_2NSiMe_2CH_2)_2$ and Ph = phenyl, which reacts with one dihydrogen molecule and forms a bridging zirconium hydride and an N−H bond. [3] Theoretical modeling predicted that this complex could dissociatively add even more than one dihydrogen molecule under appropriate experimental conditions.[4] Recently, Pool et al.[5] reported the synthesis of a dinuclear Zr complex $[(\eta^5\text{-}C_5Me_4H)_2Zr]_2(\mu^2,\eta^2,\eta^2\text{-}N_2)$, **1**, which reacts with dihydrogen at only 1 atm and 22 °C. Subsequent warming of the complex to 85 °C even leads to formation of a small amount of ammonia. Although the discovered reaction is still not catalytic, its significance and practical importance is undisputable.

Surprisingly, when a slightly different ligand is used, $C_5Me_5^-$ instead of $C_5Me_4H^-$, the reaction follows a different path with expulsion of free N_2.[6] Thus, replacing only one methyl group with hydrogen completely changes the reactivity of the complex. Very recently we explained the reason behind this remarkable difference.[7] We examined computationally a series of side-on and end-on coordination of the N_2 molecule in mono- and dinuclear complexes, $(Cp'_2Zr)(N_2)$ and $(Cp')_2Zr(N_2)Zr(Cp')_2$, where Cp′ = $C_5H_{5-n}Me_n$, n = 0−5. The results for mononuclear complexes showed that the electronic effects would favor side-on coordination, a most suitable mode for hydrogenation, when more methyl groups are added to the Cp′ ring. However, the increased number of methyl groups rapidly increases the steric repul-

* Corresponding author e-mail: dmusaev@emory.edu (D.G.M.); e-mail: morokuma@emory.edu (K.M.).

† Present address: Department of Chemistry, Universitat de les Illes Balears, 07122 Palma de Mallorca, Spain.

10.1021/ct0502561 CCC: $33.50
Published on Web 01/31/2006

sion between the monomer of the dinuclear complex and makes the end-on-coordinated complexes that are not suitable for hydrogenation more favorable.

In the present study we examine in detail structure, relative stability, and properties of the $[(\eta^5\text{-}C_5Me_4H)_2Zr]_2(\mu^2,\eta^2,\eta^2\text{-}N_2)$ complex. We investigate in detail the interplay between electronic and steric effects in this complex and examine their role in the relative stability of isomers of the complex. We analyze all possible isomers of the mononuclear $(Cp'_2Zr)(\eta^2\text{-}N_2)$ complex and several selected isomers of the dinuclear $(Cp'_2Zr)_2(\mu^2,\eta^2,\eta^2\text{-}N_2)$ complex and rationalize the role of intramolecular (both steric and electronic) interactions in relative stability of these isomers.

Methodology

To find the energetically most stable structure of $(Cp'_2Zr)_2(\mu^2,\eta^2,\eta^2\text{-}N_2)$, we need to take into account different possible orientations of the four Cp′ ligands. A complication arises from the fact that each Cp′ ring has four CH_3 groups and one H, which makes possible numerous isomers for $(Cp'_2Zr)_2(\mu^2,\eta^2,\eta^2\text{-}N_2)$ that are different from each other by the mutual orientation of the $C_5Me_4H^-$ ligands. It is very time-consuming to calculate all isomers. To simplify the task we used the following strategy. As an initial step we examined in detail all possible isomers of the mononuclear complex $(Cp'_2Zr)(\eta^2\text{-}N_2)$. Using the monomer, we hoped to select more likely monomer structures that can form stable dimers and hence reduce substantially the computational effort. As a second step, we used only the best isomers of the monomer to form several isomers of the dimer.

We used the hybrid density functional B3LYP method[8] and the Stevens-Basch-Krauss (SBK)[9] relativistic effective core potentials and the standard CEP-31G basis sets for H, C, N, and Zr atoms. Previously it was shown that d-type polarization functions on the N atoms are important for accurate prediction of the geometry and energetics of similar Zr complexes.[4b] Therefore we added a d-type polarization function for the two N atoms. Below, this approximation will be called as B3LYP/CEP-31G(d_N). All calculations were performed using the Gaussian 03 program package.[10] The atomic charges were evaluated via the Natural Population Analysis (NPA) algorithm.[11] To estimate the steric interactions, we performed molecular mechanics (MM) calculations using Universal Force Field (UFF).[12]

Results and Discussion

A. Cp'_2ZrN_2 Monomer. The structure of the monomer $(Cp'_2Zr)(\eta^2\text{-}N_2)$ taken from the experimentally determined structure of the $[(\eta^5\text{-}Cp')_2Zr]_2(\mu^2,\eta^2,\eta^2\text{-}N_2)$ dimer is shown in Figure 1(a).[5] The positions of the H atoms on Cp are indicated with arrows. As can be seen, the two H atoms are oriented in opposite directions. Figure 1(b) gives a schematic representation of all possible positions of the hydrogen atoms. The solid Cp′ ring is placed above the plane of the paper, and the dashed Cp′ ring is placed below. We distinguish different Cp′ orientations by different positions of the H atoms. Each H atom on the dashed Cp′ ring can occupy positions 2, 4, or 6, while on the solid Cp′ ring it can occupy positions 1, 3′, 3″, 5′, or 5″. Here, the larger number

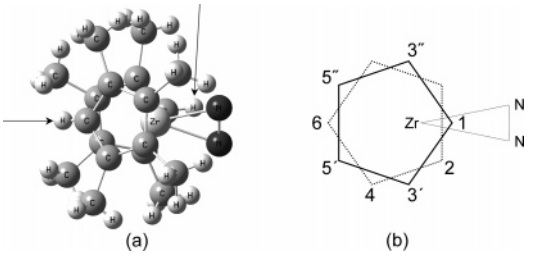


Figure 1. (a) The monomeric unit taken from the X-ray structure of $(Cp'_2Zr)_2(\mu^2,\eta^2,\eta^2\text{-}N_2)$. The positions of the H atoms are indicated with arrows. (b) Schematic representation of different H positions. 1, 2, 3, 4, 5, and 6 denote different possible positions of the H atoms. The solid Cp′ ring is placed above the plane of the paper and the dashed Cp′ ring is placed below.

corresponds to further away from dinitrogen, and with ′ and ″ indicates closer or further from position 2 or 4. We will denote isomers by the positions of their H atoms. The experimentally reported structure corresponds to isomer 16, where the first number (1) corresponds to the solid Cp′ ring, and the second number (6) corresponds to the dashed Cp′ ring.

There are a total of 13 possible isomers of the $(Cp'_2Zr)(\eta^2\text{-}N_2)$ monomer, which are shown schematically in Figure 2. We attempted to optimize all these structures. However, during the optimization some of these structures converged to other isomers. As a result, we obtained only eight different stable isomers, 56, 36, 16, 5″4, 5′2, 3′2, 12, and 3″2, presented in bold in Figure 2. The barrier for Cp′ rotation is found to be very small—only 2.7 kcal/mol for rotation from position 56 to position 36, for example. This indicates that different isomers correspond to points on a relatively flat potential energy surface.

Table 1 presents the relative energies of the eight optimized isomers. As can be seen, isomer 56 possesses the lowest energy of all. The experimentally reported structure, 16, is 0.90 kcal/mol higher in energy than isomer 56. At first sight, this finding does not agree with the experimental result. However, here we consider only the monomer and not the dimer, which was actually experimentally studied. As we will see later, when we will discuss the dimer, the calculated data fully agree with the experimental results.

The data in Table 1 represent calculations in a vacuum. The experiment of Chirik and co-workers was carried out in a pentane solution.[5] To evaluate the solvent effect, we performed single-point energy calculations of isomers 56, 16, and 12 using the PCM model.[13] The PCM relative energies of 0.0, 0.81, and 4.31 are very close to the values without solvent effects: 0.0, 0.90, and 4.50 kcal/mol, respectively. Thus, the nonpolar pentane solvent does not affect significantly the energy differences between different isomers.

Now, let us rationalize the relative energies of the eight different isomers in terms of the specific interactions within the molecule. Analyzing the NPA atomic charges,[11] we found that the C atoms of C_5Me_4H rings are not equally charged. In Figure 3 we show the charge distribution only for isomer 56, as an example. The C atoms, which are bound to CH_3

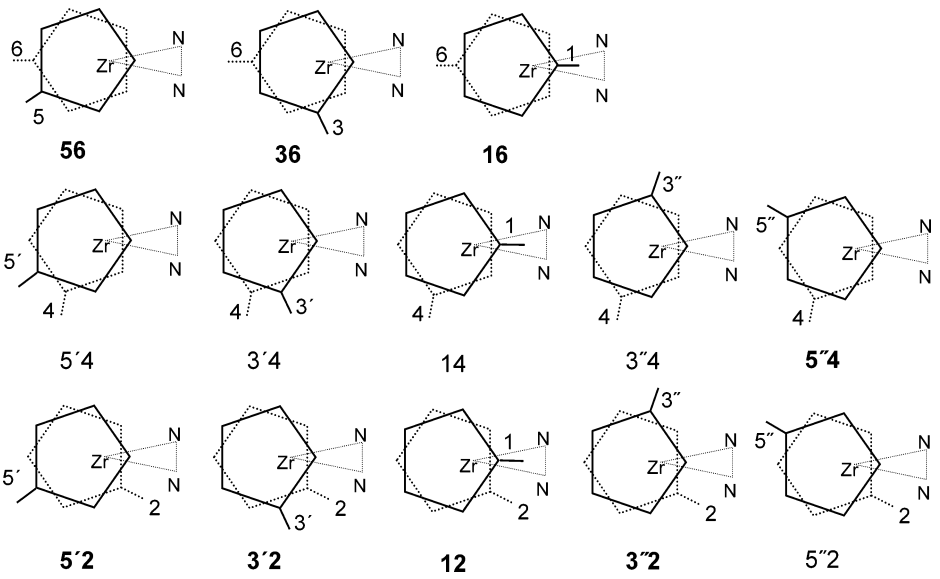


Figure 2. Schematic representation of all possible isomers of the $(Cp'_2Zr)(\eta^2\text{-}N_2)$ monomer. The stable structures are labeled in bold.

Table 1. B3LYP/CEP-31G(d_N) Calculated Relative Energies, ΔE, Dipole Moments, d, and Relative UFF//B3LYP/CEP-31G(d_N) Energies, ΔE_{steric}, of the Eight Stable Isomers of $(Cp'_2Zr)(\eta^2\text{-}N_2)$ Monomers

	ΔE (kcal/mol)	D (Debye)	ΔE_{steric} (kcal/mol)		ΔE (kcal/mol)	D (Debye)	ΔE_{steric} (kcal/mol)
56	0.00	3.22	0.00	5′2	1.81	3.79	4.74
5″4	0.49	3.26	4.50	3″2	3.07	4.17	4.18
16	0.90	3.49	5.48	3′2	3.26	4.13	4.69
36	1.02	3.68	1.45	12	4.60	4.27	4.98

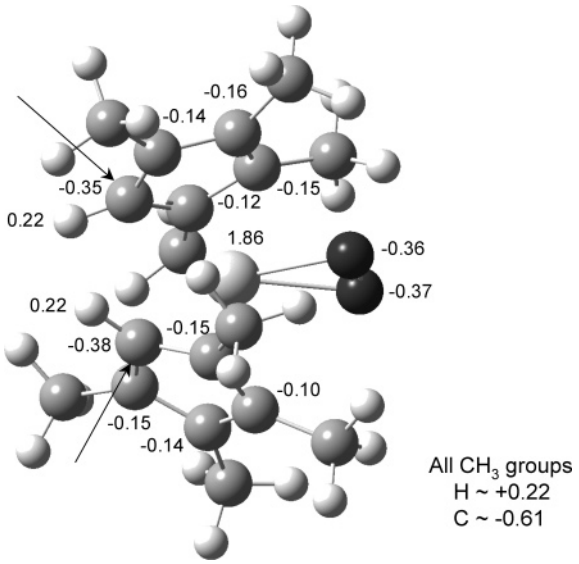


Figure 3. NPA atomic charges of isomer 56.

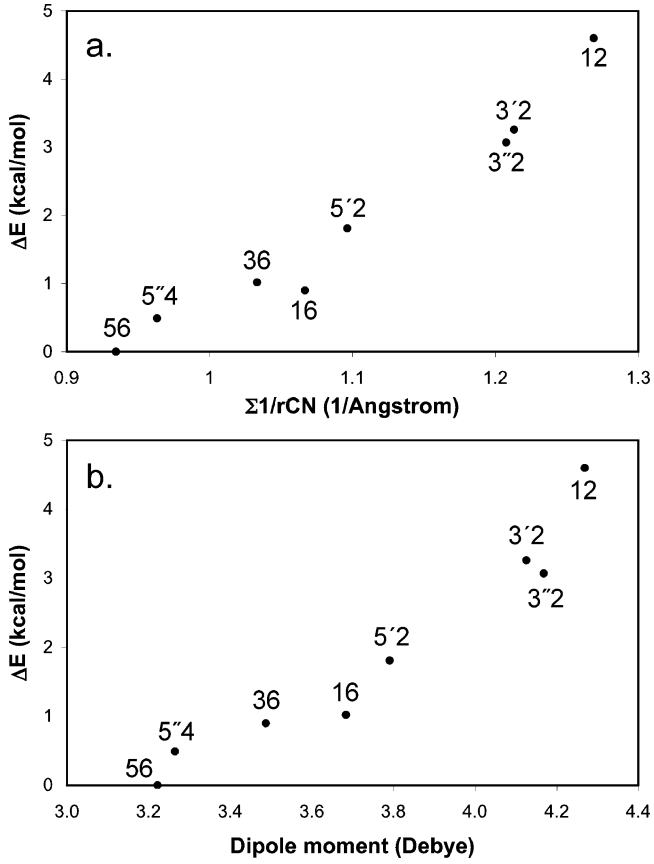


Figure 4. Dependence between relative energies of different $(Cp'_2Zr)(\eta^2\text{-}N_2)$ isomers and (a) $\Sigma 1/r_{CN}$ and (b) their dipole moments.

groups, have approximately equal charges that vary between −0.10e and −0.16e. The "special" C atom bound to the H atom (denoted as C_H) has the largest negative charge: −0.38e for the solid Cp′ ring and −0.35e for the other Cp′ ring, respectively. The charge distributions of the other possible isomers are similar to those discussed above.

We measured the distances between the two C_H atoms and the two N atoms in every isomer (r_{CN}) and plotted $\Sigma 1/r_{CN}$ against the relative energy of the isomer. The result is shown in Figure 4(a). The graphical representation shows a direct relationship between the relative energies of different isomers of Cp'_2ZrN_2 and $\Sigma 1/r_{CN}$. This indicates that the stability of the complex affected by the electrostatic repulsion between

Table 2. Solution of $\Sigma c_i N_{ij} = \Delta E_j$, for $(Cp'_2Zr)(\eta^2\text{-}N_2)$ Monomers[a]

	c_i			c_i
H−N_2	0.9520	repulsion	H−H	0.0903
Me-N_2	−0.7954	attraction	Me-H	0.0281
Me-Me	0.4036	repulsion		

[a] See text.

the negatively charged N atoms and the negatively charged C_H atoms. Another factor affecting the stability of the respective isomer is the weak electron donating CH_3 groups, which leads the charge redistribution in Cp-rings. As a result the C_H atom becomes approximately 2−3 times more negative than the other four C atoms in the C_5Me_4H ring, and, consequently, the electrostatic repulsion between the N atoms and the C_H atoms becomes larger. In the most stable isomer 56 these four atoms are located further away from each other, while in the less stable isomer 12 the two H atoms are at their closest distance from N_2 (Figure 2).

The relative stability of isomers of $(Cp'_2Zr)(\eta^2\text{-}N_2)$ could be also rationalized by counting the number of methyl−methyl, methyl−H, H−H, methyl−N_2, and H−N_2 interactions in each of them. We formed the set of linear equations $\Sigma c_i N_{ij} = \Delta E_j$, where N_{ij} is the number of interactions of type (i) methyl−methyl, methyl−H, H−H, methyl−N_2, and H−N_2, respectively, in the jth monomer and ΔE_j is its relative energy. In Table 2 we give the results of a least-squares fit for the coefficients c_i. As seen from this table, the largest contribution to the energy comes from the H−N_2 interaction, which mimics the electrostatic repulsion between the N atoms and the C_H atom. As we showed above this C_{H-N} repulsion directly affects the relative stability of the respective isomer (Figure 4(a)).

We would like to point out that in the case of C_5H_5 or C_5Me_5 ligands the five C atoms are equally charged, and the ligand has a zero dipole moment. However, in the case of the C_5Me_4H ligand, the weak electron donating methyl groups lead to a nonzero dipole moment. B3LYP/CEP-31G calculations of an individual C_5Me_4H ring show a dipole moment of 1.24 D. One may expect that the complex with C_5Me_4H ligands will be more polar and have different solubility than complexes with the nonpolar C_5H_5 or C_5Me_5 ligands.

The different orientation of the two H atoms leads to a different dipole moment of the $(Cp'_2Zr)(\eta^2\text{-}N_2)$ complexes. The fourth and eighth columns of Table 1 show the calculated dipole moments of the eight $(Cp'_2Zr)(\eta^2\text{-}N_2)$ isomers. The smallest dipole moment, 3.22 D, belongs to the most stable isomer 56. The highest-energy isomer 12 has the biggest dipole moment, 4.27 D. Interestingly, we found a direct relationship between the relative energies of the $(Cp'_2Zr)(\eta^2\text{-}N_2)$ monomers and their dipole moments (Figure 4(b)). The larger the dipole moment, the less stable is the monomer.

Calculated trends in the total dipole moment of the system can be explained by analyzing its components, the dipole moment of ZrN_2 fragment and the two dipole moments of the two Cp′ rings. The dipole moment of the individual Cp′ ring is oriented along the "special" C−H bond and points away from the H atom. The dipole moment of ZrN_2 is oriented toward the Zr atom. Thus, in the case of isomer 12 all three dipole moments add up, giving the largest dipole moment of all isomers. The two Cp′ dipoles repel from the ZrN_2 dipole, and the isomer becomes the less stable. In the case of the isomer 56 the two Cp′ dipoles are opposite to the ZrN_2 dipole, and the resulting total dipole moment is the smallest one among all eight structures. The attraction between the two Cp′ dipoles and the ZrN_2 dipole makes isomer 56 the most stable $(Cp'_2Zr)(\eta^2\text{-}N_2)$ isomer.

Table 3. Relative Energy, ΔE, Relative Electrostatic Repulsion between the N and C_H Atoms, ΔE_{ES}, and Relative Steric Repulsion, ΔE_{steric}, of Different Isomers of $(Cp'_2Zr)_2(\mu^2,\eta^2,\eta^2\text{-}N_2)$[a]

dimer	ΔE	ΔE_{ES}	ΔE_{steric}	dimer	ΔE	ΔE_{ES}	ΔE_{steric}
56−56	4.64	0.00	16.74	3″2−3″2	21.65	34.91	0
16−16	0.00	26.94	7.90	12−12	2.82	46.42	4.98
36−36	7.78	22.07	10.61	56−16	1.37	10.72	12.21

[a] All energies are in kcal/mol.

To estimate the role of steric repulsion in the relative stability of different monomers, we calculated the UFF MM energy of the B3LYP/CEP-31G(d_N) optimized structures. The results are given in the fourth and eighth columns of Table 1. As can be seen, the most stable structures 56 and 36 have the lowest steric repulsion. All other isomers have a 4.0−5.5 kcal/mol higher steric repulsion than the optimal structure 56. There is no direct relation between the steric repulsion of these structures and their relative energy. Therefore, we could conclude that in the case of $(Cp'_2Zr)(\eta^2\text{-}N_2)$ monomers, the steric repulsion does not play a critical role. The stability of the complex is determined predominantly by electronic effects (electrostatic and dipole−dipole interactions). The small barrier for the rotation of the Cp′ rings allows the realization of the most favorable orientation.

B. $(Cp'_2Zr)_2N_2$ Dimers. The actual complex that was studied experimentally is the dinuclear $[(\eta^5\text{-}C_5Me_4H)_2Zr]_2\text{-}(\mu^2,\eta^2,\eta^2\text{-}N_2)$ complex, **1**. To find the lowest-energy structure of this complex, it would be best if we could consider all possible orientations of the four Cp′ rings. However, the number of possible isomers of the dimer is very large, and searching for all these is not practical. Therefore we decided to limit our study to the eight candidates, which are formed by two identical monomers, 56−56, 5″4-5″4, 16−16, 36−36, 5′2−5′2, 3″2-3″2, 3′2−3′2, and 12−12, where the dimer notation combines those of the two monomers. We considered only one "mixed" dimer, 56−16, formed by the lowest-energy monomer 56 and monomer 16, which corresponds to the experimental structure. We attempted to optimize all nine structures, but during the optimization some of them converged to other isomers. As a result, we obtained six different stable dimers, 56−56, 16−16, 36−36, 3″2-3″2, 12−12, and 56−16.

The relative energies of the investigated dimers are given in Table 3. As can be seen, isomer 16−16 has the lowest energy among all examined possibilities. This is in excellent agreement with the experimentally found structure of **1** (Figure 1a). Table 3 gives also the relative electrostatic repulsion between the two negatively charged N atoms and

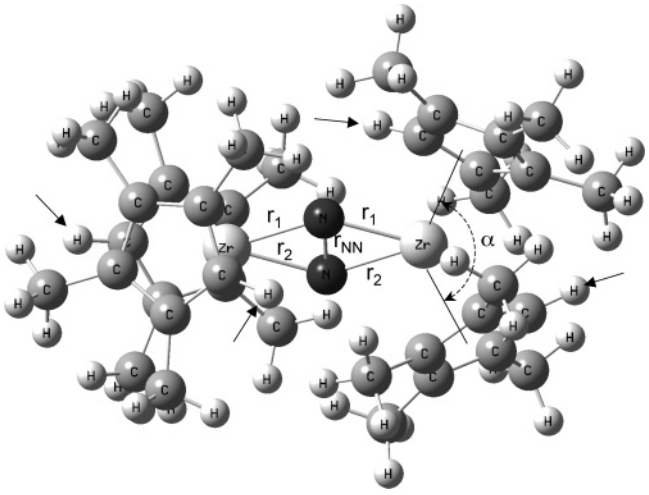


Figure 5. Optimized structure of the 16−16 isomer of (Cp'_2-Zr)(η^2-N_2). The positions of the H atoms are indicated with arrows.

the four negatively charged C_H atoms, ΔE_{ES}, calculated in analogy with the monomer case. As could be expected, the electrostatic interaction favors the 56−56 dimer, formed by the two optimal monomers 56. The 56−56 dimer has the lowest electrostatic repulsion among all investigated dimers. However, the lowest energy dimer 16−16 is not formed by two 56 units but by two 16 units. This is due to the fact that, while not crucial for the monomer case, the steric interactions play an important role in the dimer. In monomer 56 the two H atoms point away from the dinitrogen molecules. Thus, when two 56 units are combined together, four H atoms point away from N_2. The two 56 units "see" each other via their methyl groups, which leads to higher steric repulsion than that in the 16−16 dimer. In the 16−16 dimer, two of the H atoms point toward the N_2 molecule and thus reduce the steric repulsion. We estimated the steric effect by performing the UFF MM calculation at the B3LYP/CEP-31G(d_N) optimized geometry of each dimer. The results are shown in the fourth and eighth columns of Table 3. Indeed, isomer 16−16 has 8.84 kcal/mol lower steric repulsion than the electrostatically favored 56−56 dimer. On the other hand, steric interactions do not alone govern the stability of the dimer. Isomers 3″2-3″2 and 12−12 have the lowest steric repulsion, but they are 21.65 and 2.82 kcal/mol, respectively, less stable than the lowest-energy dimer 16−16. Thus one can conclude that the relative stability of the dimers is determined by the competition between the electrostatic effects and the steric repulsion.

Figure 5 presents the optimized structure of the most stable dimer 16−16. Our calculations show that this structure does not have an imaginary frequency. Table 4 compares some key geometrical parameters of structure 16−16 with the experimental results, with the definition of the parameters given in Figure 5. Calculations predict a slightly shorter Zr−N bond and a slightly longer N−N bond than the experiment. The calculated bite angle, $\alpha_{Cp'-Zr-Cp'}$, is very close to the experimental value. In general, calculated geometries are in very good agreement with their experimental values.

Table 4. Comparison of Experimental and B3LYP/CEP-31G(d_N) Optimized Geometrical Parameters of Isomer 16−16 of (Cp'_2Zr)$_2$(μ^2,η^2,η^2-N_2) Complex, **1**[a]

	exptl[5]	calcd		exptl[5]	calcd
r_1	2.131	2.101	r_{NN}	1.377	1.404
r_2	2.119	2.114	$\alpha_{Cp'-Zr-Cp'}$	128.7	128.4

[a] Bond lengths in Å, angles in degrees. For definition of parameters see Figure 5.

Conclusions

We have demonstrated that the stability of (Cp'_2Zr)$_2$(μ^2,η^2,η^2-N_2), Cp′ = η^5-C_5Me_4H, depends on competing electronic and steric factors. Examining all possible isomers of the (Cp'_2Zr)(η^2-N_2) monomers, we have found a direct relationship between their relative energy and the electrostatic repulsion between the negatively charged N atoms of the N_2 molecule and the negatively charged C_H atoms of the Cp′ ring that are bound to H. Substitution of the H ligand by methyl groups significantly changes the charge distribution in cyclopentadienyl rings and increases the negative charge of the C_H atom and hence the electrostatic repulsion between the N atoms and the C_H atoms. As a result, the H to CH_3 substitution significantly affects the relative stability of the studied isomers. This observation indicates that it would not be appropriate to use models where the methyl groups in cyclopentadienyl rings are treated with methods that do not account for electronic effects, like Molecular Mechanics. When two monomer units are combined together to form a dimer, the significance of steric factors increases and the minimal-energy structure is determined by a combination of electronic and steric factors.

Acknowledgment. We acknowledge Prof. P. J. Chirik for providing us with the experimental X-ray coordinates. D.Q. thanks the Ministerio de Educación y Ciencia of Spain for postdoctoral support. This work was supported in part by grant (CHE-0209660) from the U.S. National Science Foundation. Computer resources were provided in part by the Air Force Office of Scientific Research DURIP grant (FA9550-04-1-0321) as well as by the Cherry Emerson Center for Scientific Computation at Emory University.

Supporting Information Available: Cartesian coordinates of the lowest-energy structure of [(η^5-C_5Me_4H)$_2$Zr]$_2$-(μ^2,η^2,η^2-N_2), **1** (isomer 16−16) (Table S.1). This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) (a) Fryzuk, M. D., Johnson, S. A. *Coord. Chem. Rev.* **2000** *200*−*202*, 379−409. (b) *Catalytic Ammonia Synthesis*; Jennings, J. R., Ed.; Plenum: New York, 1991. (c) Fryzuk, M. D. *Nature* **2004**, *427*, 498−499. (d) Shaver, M. P.; Fryzuk, M. D. *Adv. Synth. Catal.* **2003**, *345*, 1061. (e) Schlögl, R. *Angew. Chem., Int. Ed.* **2003** *42*, 2004−2008. (f) Musaev, D. G. *J. Phys. Chem. B* **2004** *108*, 10012−10018. (g) Mori, M. *J. Organomet. Chem.* **2004**, *689*, 4210−4227.

(2) *Encyclopedia Britannica*; Encyclopedia Britannica Inc.: 1997.

(3) Fryzuk, M. D.; Love, J. B.; S. J. Retting, S. J. *Science* **1997**, *275*, 1445−1447.

(4) (a) Basch, H.; Musaev, D. G.; Morokuma, K. *Organometallics* **2000**, *19*, 3393−3403. (b) Basch, H.; Musaev, D. G.; Morokuma, K. *J. Am. Chem. Soc.* **1999**, *121*, 5754−5761.

(5) Pool, J. A.; Lobkovsky, E.; Chirik, P. J. *Nature* **2004**, *427*, 527−529.

(6) Manriquez, J. M.; Bercaw, J. E. *J. Am. Chem. Soc.* **1974**, *96*, 6229−6230.

(7) Bobadova-Parvanova, P.; Wang, Q.; Morokuma, K.; Musaev D. G. *Angew. Chem. Int. Ed.* **2005**, *44*, 7101−7103.

(8) (a) Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098−3100. (b) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785−789. (c) Becke, A. D. *J. Chem. Phys*. **1993**, *98*, 5648−5652.

(9) (a) Stevens, W. J.; Basch, H.; Krauss, M. *J. Chem. Phys.* **1984**, *81*, 6026−6033. (b) Stevens, W. J.; Krauss, M.; Basch, H.; Jasien, P. G. *Can. J. Chem*. **1992**, *70*, 612−615. (c) Cundari, T. R.; Stevens, W. J. *J. Chem. Phys.* **1993**, *98*, 5555−5565.

(10) *Gaussian 03, Revision C.01*; Frisch, M. J., et al. Gaussian, Inc.: Wallingford, CT, 2004.

(11) Reed, A. E.; Curtiss, L. A.; Weinhold, F. *Chem. Rev.* **1988**, *88*, 899−926.

(12) Rappé, A. K.; Casewit, C. J.; Colwell, K. S.; Goddard, III, W. A.; Skiff, W. M. *J. Am. Chem. Soc.* **1992**, *114*, 10024−10035.

(13) (a) Miertus, S.; Scrocco, E.; Tomasi, J. *Chem. Phys.* **1981**, *55*, 117. (b) Mennucci, B.; Tomasi, J. *J. Chem. Phys.* **1997**, *106*, 5151−5158.

CT0502561

JCTC Journal of Chemical Theory and Computation

Conformational Studies of Polyprolines

Haizhen Zhong* and Heather A. Carlson

Department of Medicinal Chemistry, College of Pharmacy, The University of Michigan, 428 Church Street, Ann Arbor, Michigan 48109-1065

Received July 27, 2005

Abstract: Proline rich peptide sequences are very important recognition elements that have a significant bias toward the *all-trans*-polyproline type II (P_{II}) conformation. Our gas-phase quantum mechanics calculations at the B3LYP/6-31G* level of theory are in good agreement with previous experimental and theoretical studies. They show that *all-trans*-proline conformations are energetically more favorable than *all-cis*-polyprolines (P_I, polyproline type I). Estimates of the solvent effects show that the condensed phase can make the P_I form more populated in the correct environment. Our survey of proline oligomers in the Protein Data Bank confirmed that the predominant conformations from our calculations are seen experimentally. More importantly, we propose two new secondary structures for polyprolines, namely polyproline type III and type IV (P_{III} and P_{IV}). P_{III} is a right-handed, "square helix" from *trans*-proline oligomers. P_{IV} is a β-sheet form of *cis*-prolines. As suggested by its calculated IR spectra, the P_{III} form shares characteristics of both the P_I and P_{II} forms: it has *trans*-amide rotamers similar to P_{II} and forms a right-handed helix like P_I. We propose that the high-energy P_{III} form could exist as a conformational intermediate between P_I and P_{II}. These new forms also show that the handedness of polyproline helices depends not only on the peptide rotamers (cis or trans) but also on the values of the ψ torsions. Changing the ψ torsion from approximately 140° to approximately −30° causes the trans oligomers to flip from a typical left-handed P_{II} to a right-handed helix. Likewise, as the ψ torsion of the *cis*-proline oligomers changes from roughly 165° to −30°, the conformation changes from a characteristic right-handed P_I to a β-sheet.

Introduction

Recognition of proline-rich sequences plays a pivotal role in protein−protein interaction. The most common conformation of these sequences is the polyproline type-II (P_{II}) helix, a left-handed helix consisting of *trans*-prolines with three residues per turn (designated as 3_1 helix), $\phi \sim -75°$, and $\psi \sim 145°$. Another well-studied conformation of polyproline is the polyproline type-I (P_I), a right-handed helix with *cis*-prolines at 3.3 residues per turn (a 10_3 helix), $\phi \sim -75°$, and $\psi \sim 165°$. The ϕ and ψ angles for the P_I and P_{II} helices fall within the allowed β-strands region on the Ramachandran plot, due to the unique property of the imino proline linkage. Peptides adopting the P_{II} conformation have the propensity to function as recognition elements that bind to proline recognition domains, such as Src homology (SH3)[1], WW (named after a conserved Trp-Trp motif),[2] Enabled/VASP homology domains (EVH1),[3] the Gly-Tyr-Phe (GYF) domains,[4,5] and profilin proteins.[6,7] Recent findings have revealed a significant bias toward the P_{II} conformation in unfolded peptides and thus is a dominant component of the denatured states of proteins.[8−10]

Given the importance of proline-rich motifs, a full understanding of the P_{II} and P_I conformations of these species is highly desirable. A large body of experimental data has been reported since the crystallization of polyproline II[11] and polyproline I[12] four decades ago. However, some experiments

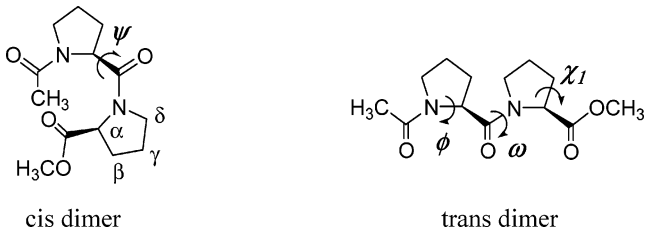
* Corresponding author phone: (336)334-5121; fax: (336)334-5402; e-mail: h_zhong@uncg.edu. Current address: Center for Drug Design, Department of Chemistry and Biochemistry, University of North Carolina at Greensboro, Greensboro, NC 27402.

10.1021/ct050182t CCC: $33.50
Published on Web 02/02/2006

provide conflicting results. For example, Chao and Bersohn observed that in aqueous solutions the proline oligomer $^{+}H_2$-NPro-$(Pro)_n$-CO_2^- predominantly adopted the P_{II} conformation;[13] the conformations of a series of proline oligomers (*tert*-butyloxycarbonyl-L-Pro_n benzyl esters, Boc-$(Pro)_n$-OBn, $n = 2-6$) in chloroform are found to exist in nearly equal populations of *cis*- and *trans*-proline conformations when $n = 2$, 3, 4 and that they abruptly assume an *all-trans*, P_{II} helical structure when n is greater than 5.[14] However, Zhang and Madalengoitia have found that the ^{1}H NMR spectra of Boc-$(Pro)_n$-OBn (n=2−5) in $CDCl_3$ suggest the *trans*-amide bond is the predominant conformation.[15] Using the $[Pro_n+H]^+$ (n=5−11) model system, Counterman and Clemmer[16] discovered that the *all-cis*, P_I helix is favored and the helix adopted an extended form while the trans adopted a compact form, contrary to previous studies. The preference of P_I helix in Counterman and Clemmer's ionized model is attributed to the N-terminus cation that stabilizes the *cis*-proline. However, this type of ionized proline, capable of forming internal hydrogen bonds, is not found in proteins containing polyproline oligomers.

Many ab initio quantum mechanics calculations on proline motifs have been carried out on proline derivatives, such as *N*-acetyl-*N*′-methylprolineamide (Ac-Pro-NHMe),[17−20] *N*-formyl-L-prolineamide (For-L-Pro-NH_2),[21] proline dimer (For-L-Pro-L-Pro-NH_2),[22] and the neutral form of proline.[23,24] These proline monomer/dimer systems are capable of forming an internal hydrogen bond, increasing the percentage of the cis conformation in both the gas phase and less polar solvents. Tanaka and Scheraga found that trans conformers of the L-proline oligomers, Ac-$(Pro)_n$-OMe where $n = 2-5$, have the lowest energy based on a simplified molecular mechanics energy function.[25] Bour et al. carried out ab initio calculations using the SV and SV(P) split valence basis sets on the proline oligomer Ac-$(L\text{-}Pro)_9$-NHMe for both the P_I and P_{II} helical conformations and found that the relative energy of the optimized P_{II} is 3.3 kcal/mol below that of the P_I.[26] However, the starting structures for optimizing P_I and P_{II} in Bour's study were generated from idealized P_I and P_{II} helical structures and therefore failed to consider the influence of the ring puckering. A statistical survey of nonredundant X-ray protein chains from the 2000 version of PDB by Vitagliano et al. observed a correlation between proline puckering and peptide bond conformation.[27] Of the 178 *cis*-proline residues in these structures, 81% adopt a downward puckered conformation. However, for the *trans*-proline residues, both upward and downward pucker conformations are observed with equal frequency. A survey of the HOMSTRAD database for the P_{II} helices revealed that although P_{II} helices only represent 3% of the residues in the database, about 60% of the proteins chains contain one or more P_{II} helices.[28] The P_{II} helices in the data set are defined as a set of sequences that have the characteristic ϕ and ψ torsions.

Here, we present a survey of the Protein Data Bank (PDB) where we found that proline oligomers are only present in all trans conformations for $(Pro)_n$ ($n \geq 5$). The trans conformation again predominates for $n = 4$, except for 3 cases which have a *cis*-proline at the beginning of the sequence. To understand the preference of *trans*-proline in proteins, we undertook calculations of simple, model polyproline oligomers from monomers to hexamers. The conformational analyses of different proline oligomers, based on ring puckering and peptide rotamers, are also discussed.



cis dimer trans dimer

Figure 1. The structures of *cis*- and *trans*-L-proline dimer. Cα through Cδ are labeled, and the ψ, ϕ, ω, and χ_1 torsions are noted.

The distribution and conformational behavior of proline oligomers were examined through a series of ab initio RHF/6-31G* and B3LYP/6-31G* calculations. Proline hexamers were constructed using the structures of proline dimers (Figure 1), which were derived from the conformational scans of dimers and monomers at the RHF/6-31G* level. Figure 1 shows the cis and trans conformers of the L-proline dimer. All conformers in the calculations include the peptide rotamer (cis/trans, as 'c' or 't') and pyrrolidine ring puckering (up/down, as 'U' or 'D') conformation. The calculations on the monomer, dimers, and hexamers of L-proline suggest that the trans conformation has the lowest energy. This finding supports our observation in the PDB that *trans*-proline is the dominant conformation in polyprolines. In addition to revealing the ring puckering effect and the solvation effect on the P_I and P_{II} conformers, we also propose two new regular secondary structures of polyproline, herein identified as polyproline type-III (P_{III}) and polyproline type-IV (P_{IV}).

Methods

Survey of Proline Oligomers in the PDB. The July 2003 release of the PDB was searched for polyproline sequences. The classification of oligomers Pro_n is based on the number of proline residues in a contiguous sequence. The dihedral angles ω, ϕ, ψ, and χ_1 were measured for all entries where $n \geq 4$ (Figure 1). As shown in Figure 1, the ω torsion is defined as the dihedral among [$C\alpha_i$-C_i-$N_{(i+1)}$-$C\alpha_{(i+1)}$]; the ϕ torsion as [C_i-$N_{(i+1)}$-$C\alpha_{(i+1)}$-$C_{(i+1)}$]; the ψ torsion as [$N_{(i)}$-$C\alpha_{(i)}$-$C_{(i)}$-$N_{(i+1)}$]; and the χ_1 torsion as [N_i-$C\alpha_i$-$C\beta_i$-$C\gamma_i$]. The torsion ω defines the proline amide bond as cis ($\omega = 0°$) or trans ($\omega = 180°$, abbreviated herein as 'c' and 't' for cis and trans, respectively). The torsional angles ϕ and ψ determine the secondary structure of polyproline oligomers. The torsion χ_1 describes the pucker conformation of the proline ring, where $\chi_1 > 0$ denotes "endo" or "pucker-down" and $\chi_1 < 0$ is "exo" or "pucker-up" (abbreviated as 'D' and 'U', respectively).

Computational Methods. All ab initio calculations were carried out using Gaussian98 and Gaussian03.[29] In all cases, the default convergence criteria were used. Systematic conformational scans were carried out at the RHF/6-31G* level of theory for the monomers and dimers to characterize the minima to be used for constructing the hexamers. All reported minima along the potential energy surface (PES)

Table 1. Distribution of $(Pro)_n$ $n = 4$ Conformations (50 Entries), Based on the Proline Amide Bond (t/c) and the Ring Puckering (D/U)

conformation	no. of entries	percentage (%)	conformation	no. of entries	percentage (%)
tDtUtUtU	9	15.5	tDtDtUtD	3	5.2
tDtDtDtD	8	13.8	tDtUtDtD	3	5.2
tDtUtDtU	7	12.1	tDtDtUtU	2	3.4
tUtDtUtU	6	10.3	cDtDtUtU	2	3.4
tDtUtUtD	6	10.3	tUtDtUtD	1	1.7
tDtDtDtU	5	8.6	tUtUtDtU	1	1.7
tUtUtUtU	4	6.9	cDtDtUtD	1	1.7

were subject to full geometry optimizations and were further confirmed through frequency calculations, which gave no imaginary frequencies.

Full conformational sampling of eight different dimers (i.e., tDtD, tDtU, tUtD, tUtU, cDcD, cUcU, cDcU, and cUcD) was carried out by scanning the ψ torsion for a full 360° at 10° increments. The optimized minima of the dimers were used to construct the initial conformations of the hexamers; full geometry optimizations of these hexamers were carried out without any constraints at RHF/6-31G*. The resulting RHF/6-31G* minima were used as the starting point for the B3LYP/6-31G* optimization. For the RHF/6-31G* minimizations, the optimized hexamers were verified through frequency analysis and were visualized using XChemEdit.[30] Frequency calculations for the B3LYP/6-31G* minima were not possible because of extensive memory requirements, but the structures were very similar to those obtained with RHF/6-31G*. Solvent effects were included for the hexamers by performing single-point energy calculations using self-consistent reaction field (SCRF) theory with the isodensity surface polarized continuum model (IPCM) at the RHF/6-31G* level.[31] The SCRF-IPCM calculations were carried out for solvent dielectric constants of 4.90, 32.63, and 78.39, corresponding to chloroform, methanol, and water, respectively. A combination of 44 phi points and 22 theta points (parameters for the radial grid employed in IPCM) was adopted for the SCRF-IPCM calculations.

Results and Discussion

Distributions of Polyprolines in the Protein Data Bank. Our survey of the PDB showed that more than 6300 of the 22 119 PDB entries (28.5%) contain at least one proline dimer in their sequence; 475 entries (2.1%) contain at least one proline trimer. Some proteins contain nine consecutive proline residues (e.g., farnesyltransferase[32]) and 15 consecutive prolines (e.g., profilin[7]). Downward ring puckers were seen slightly more often, but both up and down ring puckers occurred in large numbers. This implies that the two are nearly equal in free energy with the downward rings being slightly more favorable. It was interesting that no *cis*-prolines were observed in oligomers with five or more consecutive prolines and only 3 of the 58 tetramers were found to contain a *cis*-proline. These *cis*-prolines only occurred at the beginning of the tetramer motifs (Table 1).

Calculations of Proline Monomers. Systematic searching of the conformational space for the L-proline monomer (Ac-Pro-OMe) yielded 8 minima, two for each of the four combinations (i.e., tD, cD, tU, cU) examined. The characteristic torsion angles (ϕ and ψ) and energies are given in Table 2. The PES for scanning ψ (actually, the ester torsion N-Cα-C-OMe) in all four conformers noted above are provided in the Supporting Information (Figures 1S−2S). The energy differences among the eight minima are less than 3.00 kcal/mol [with the exception of cU(−30)]. The trans conformers are energetically more favorable than the cis monomers. For the same rotamers, the endo ('D') conformation is slightly more favorable in energy than the exo ('U'). The global minimum is tD ($\phi = -70°$, $\psi = 157°$). This finding agrees very well with studies from other groups.[18,25,33] Our calculations on the proline monomer agree well with experimental data showing that the endo conformation is the most populated conformation in the model molecule H_2N-Pro-COOH in the gas phase.[34] Table 2 shows that ϕ torsions for pucker-up conformers are about −60°, while the pucker-down conformers have a ϕ torsion around −70°. The conformers with ester torsions in the range of 150°−170° are lower in energy than those with ester torsions in the range of −30° to −20°. The difference in energy and the conformation can be explained by the 1−4 distance between the two carbonyl carbons. The distance is slightly greater in the 'D' conformation, compared to the 'U' conformation, to effectively attenuate the steric hindrance. This observation was also noted by Vitagliano et al.[27] The larger ψ torsions in the cis conformations result from alleviating the steric interactions between the N-terminal methyl group and the C-terminal methoxyl group. In the dimer and hexamer studies, the minima yielded ester torsions between 150°−170°. Additional minima with different ester torsions were not pursued.

Table 2. Energies and Characteristics of Minima for the Proline Monomer, Ac-Pro-OMe

conformer	characteristics[a]	ϕ (°)	ψ (°)	ΔE (kcal/mol)
1	tD(160)	−70.0	157.0	0.00
2	tD(−20)	−67.4	−23.5	1.21
3	tU(150)	−59.5	148.4	0.97
4	tU(−30)	−55.2	−34.3	1.45
5	cD(170)	−75.4	171.0	1.74
6	cD(−20)	−75.9	−16.5	2.48
7	cU(170)	−62.5	167.7	2.83
8	cU(−30)	−60.2	−34.5	3.20

[a] The characteristics note the peptide rotamer, ring pucker, and ester torsion (ψ for monomers).

Calculations for Proline Dimers. The monomer conformations with the ψ torsion in the range of 150°−170° are lower in energy, so they were used to initiate calculations of the dimers (Ac-Pro-Pro-OMe). The torsional scans for dimers gave two minima in most cases (Figure 2 and Table 3), one with ψ torsions in the range of 130°−170° and the other with range of −50° to 10°. In the most favorable states, the range of ψ torsions for the trans dimers is 130°−150°, while that range for the cis dimers is 160°−170°. Among the eight low-energy conformers, the trans states are more energetically favorable than the cis states. The cis dimers are at least 4 kcal/mol higher in energy than their trans counterparts. This finding is consistent with studies from

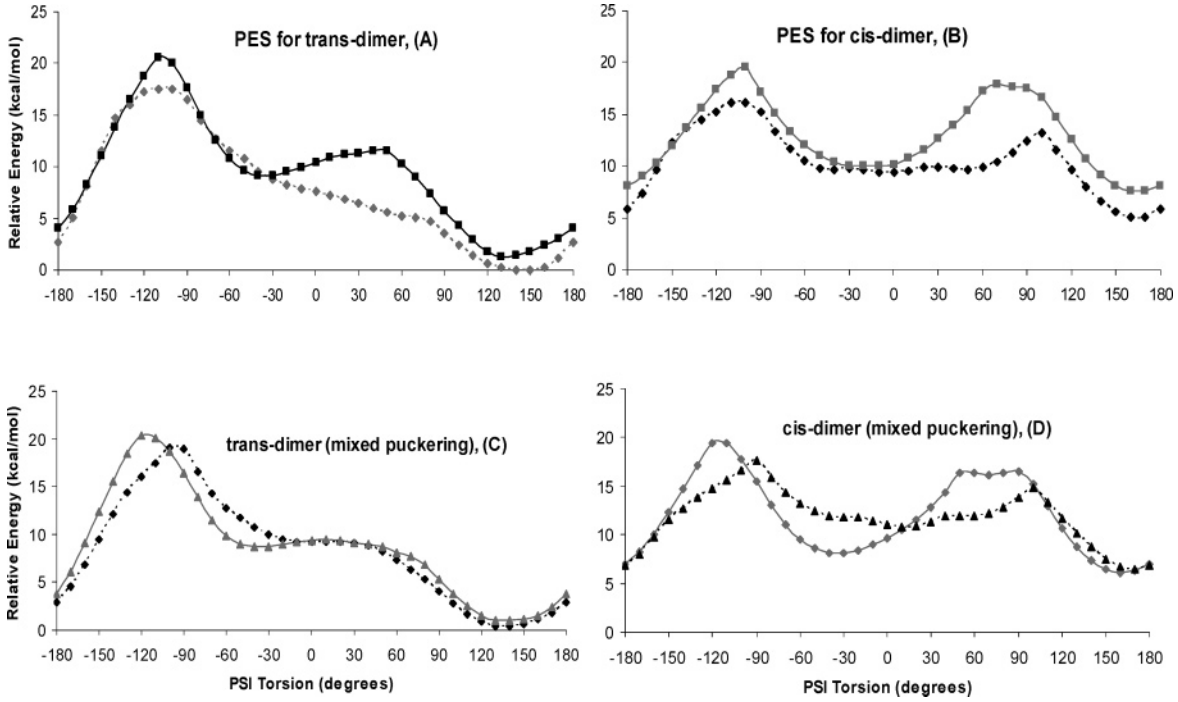


Figure 2. PES for various dimers. All energies are relative to the lowest energy point in the scans, tDtD(148). (A): black square: tUtU; gray diamond: tDtD; (B): gray square: cUcU; black diamond: cDcD; (C): gray triangle: tUtD; black diamond: tDtU; (D): gray diamond: cUcD; black triangle: cDcU.

Table 3. Energies and Characteristics of the Minima for the Proline Dimer (Ac-Pro-Pro-OMe)

confor-mers	charac-teristics[a]	ϕ_1 (°)	ψ (°)	ϕ_2 (°)	ΔE (kcal/mol)	dipole (Debye)
1	tDtD(148)	−70.0	148.2	−74.5	0.00	2.19
2	tDtU(135)	−69.9	135.8	−63.5	0.32	1.29
3	tUtD(135)	−60.8	135.7	−78.8	1.02	1.83
4	tUtU(133)	−60.6	133.4	−64.3	1.30	1.81
5	cDcD(165)	−73.1	165.2	−75.8	4.98	9.11
6	cUcD(161)	−57.2	161.4	−78.3	6.09	9.13
7	cDcU(169)	−73.3	169.0	−65.7	6.43	9.24
8	cUcU(165)	−57.8	165.5	−67.7	7.55	9.35
9	cUcD(−34)	−67.4	−34.2	−83.3	8.10	3.27
10	tUtD(−38)	−66.5	−38.6	−76.5	8.70	6.43
11	tUtU(−35)	−64.9	−35.5	−60.1	9.11	6.71
12	tDtU(4)	−89.8	4.0	−55.8	9.17	6.49
13	cDcD(−3)	−88.6	−3.2	−79.5	9.39	2.47
14	cDcD(−40)	−69.9	−40.5	−82.7	9.64	4.20
15	cDcD(49)	−138.6	49.1	−86.2	9.67	3.82
16	cUcU(−12)	−75.7	−12.8	−59.3	9.96	2.76
17	cDcU(13)	−92.4	13.8	−55.3	10.83	2.51
18	cDcU(56)	−140.4	56.1	−77.3	11.90	4.17
19	cUcD(68)	43.8	68.6	−85.8	16.17	5.32

[a] The characteristics note the peptide rotamer, ring pucker, and ψ torsion. Ester torsions (N_2-Cα_2-C_2-OMe) were between 140° and 175° for all conformers.

other groups.[25] For the same peptide rotamers ('t' or 'c'), pucker-down conformations are slightly lower in energy than the pucker-up conformations. The eight lowest-energy conformers of the dimers are shown in Figure 3, and the higher-energy conformers are shown in Figure 4. The eight low-energy conformers have more favorable steric interactions between the two pyrrolidine rings. The proximity of both termini in the cis dimers (Figure 3) causes them to be higher in energy than the trans dimers. The energetic difference between cis and trans dimers may explain the absence of any *all-cis*-polyprolines ($n \geq 4$) in the current PDB.

For the low-energy trans dimers, the energy difference between ring puckers is small (less than 1.30 kcal/mol), indicating that these conformers would be nearly equally populated at room temperature. Quan and Wu's calculations for two triple helices at HF/6-31G* level suggest that different puckering modes have very similar energies.[35] Because of the closeness in energy between tDtD, tDtU, tUtD, tUtU (Table 3), it is likely that any combination of the above four dimers would give rise to an energetically feasible polyproline oligomer. This hypothesis is consistent with the proline oligomer distribution in the PDB, where the five most common tetramer motifs are comprised of these low-energy dimer moieties (Table 1). For the low-energy dimers with cis rotamers, the downward puckered conformation is energetically more favorable than the upward puckered one. These observations are in good agreement with the survey by Vitagliano et al.[27]

For the proline dimers with ψ torsions in the range of 130°−150°, trans conformers (1−4 in Table 3) have ϕ torsions between −80° and −60°. Therefore, according to the definition of P_I and P_{II} helices, they adopt a P_{II} helix. The cis conformers (5−8 in Table 3) possess a P_I helix, with ϕ torsions in the range of −80° to −55° and ψ torsions of 160°−170°. This finding supports the observations by Zhang and Madelengoitia that dimers can have characteristic values of ϕ and ψ which correspond to P_I and P_{II} helices.[15]

Conformations 9−11 and 14 in Table 3 show that another common minimum conformation includes ψ near −40°, and they are several kcal/mol higher in energy than the lowest global minimum. However, the conformations for tDtD and

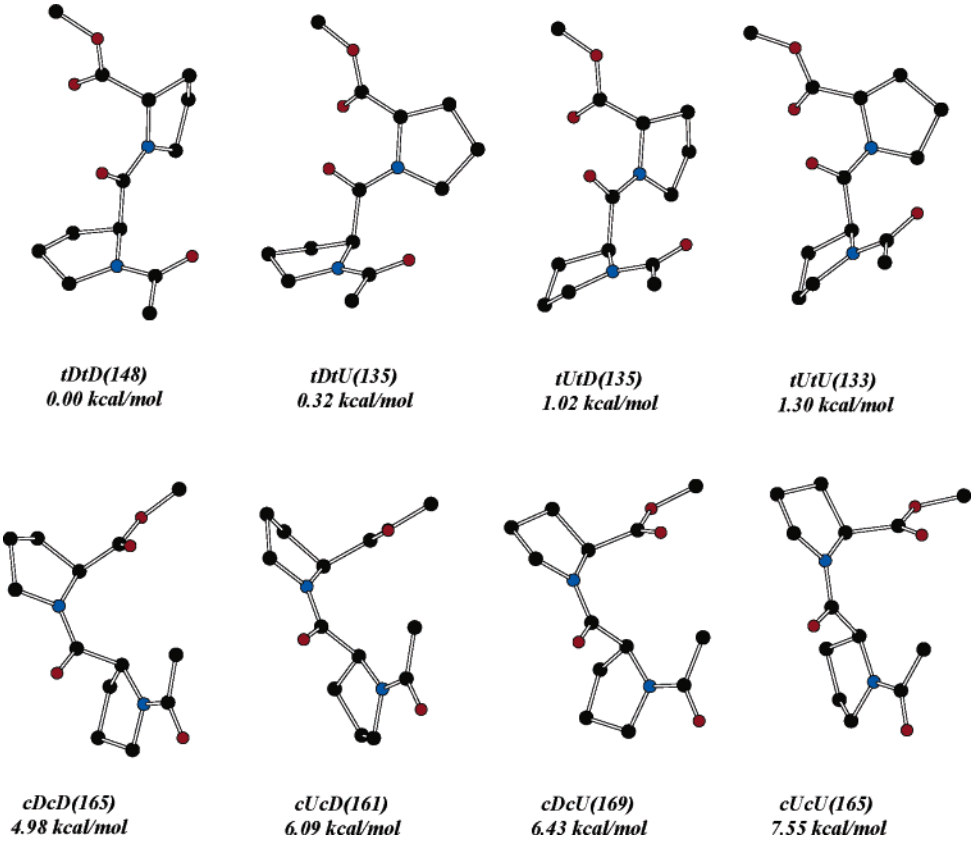


Figure 3. Low-energy minima of dimers illustrating various rotamers and puckering conformations. Hydrogen atoms are not shown for clarity. Color codes for atoms: black: C; red: O; blue: N.

tDtU at $\psi = -40°$ lie along a shoulder on the PES (Figure 2A,C). Full optimization of the tDtU conformer gave a minimum with a ψ torsion of 4.0°. This minimum was verified by the frequency calculations. Optimization of the tDtD conformer failed to yield a minimum outside the range of 130°−170°. A similar situation occurs with cUcU and cUcD, where cUcD gave rise to a minimum with ψ near 70°, while cUcU failed to yield a minimum with ψ near 70° (Figure 2B,D). Interestingly, three energetically close minima for cDcD and cDcU exist (Figure 2D) with ψ torsions in the range of −60° to 80°. Full optimization and frequency calculations reveal that five of these six stationary points yield unique minima. Full optimization of the cDcU conformation ($\psi = -40$) caused an interconversion to the cDcD conformer [designated as cDcD(−40)]. A high-energy conformation like cDcU(−40) most likely has a very small barrier to rearrangement to the lower energy cDcD(−40), and it appears that it is not a stable minimum. Ring inversions such as this one are not uncommon for prolines. Badoni et al.[36] reported that the barrier for ring inversion from the 'U' conformation to the 'D' is 2.1 kcal/mol for *N*-formyl-*trans*-proline amide (For-Pro-NH_2, a proline monomer) based on B3LYP/6-31G* calculations. Kang and Park[37] also reported an estimated energy barrier of 2.2 kcal/mol for ring inversion from the 'D' conformation to the 'U' conformation for *N*-acetyl-L-proline-*N*′ and *N*′-dimethylamide (Ac-Pro-NMe2) at the B3LYP/6-311++G** levels.

The high energy of trans dimers with ψ torsions near −40° and 0° (Figure 4) arises from strong repulsion between the atoms of the N-terminal pyrrolidine ring and the methylene group at the δ-position of the second proline residue. For the cis dimers with similar ψ torsions, the high energy is due to the steric overlap between the first pyrrolidine ring and the C-terminal ester group and/or the steric interactions from both termini.

One new and interesting observation from the torsional scan of the dimers is that the shape of the potential energy surface seems to be determined by the peptide bond conformation and ring puckering of the first proline (Figure 2). The PES of tDtU is more similar to tDtD than tUtU. Similarly, tUtD resembles tUtU, cDcU resembles cDcD, and cUcD resembles cUcU.

Calculations for Proline Hexamers. The energetic penalty of pushing any dimer conformation away from its most favorable ψ torsion of ∼160° is less than 10 kcal/mol (Figure 2) and comparable to the results of Mattice et al.[38] Therefore, we chose to include these higher-energy conformations in our analyses of polyproline hexamers. The proline hexamers were constructed using characteristics of the optimized dimers. We did not pursue conformations of the hexamer based on conformers 15, 18, or 19 in Table 3; propagating helixes with these ϕ and ψ combinations resulted in oligomers that collided back upon themselves in an unrealistic fashion. It should also be noted that helices based on the dimer conformers 13 and 14 in Table 3 minimized to the same conformation (conformation 14 in Table 4). This resulted in a total of 15 conformations for the hexamers.

Each appropriate conformer of the hexamer was fully optimized using the RHF/6-31G* levels of theory, and the resulting minima were further optimized using B3LYP/6-

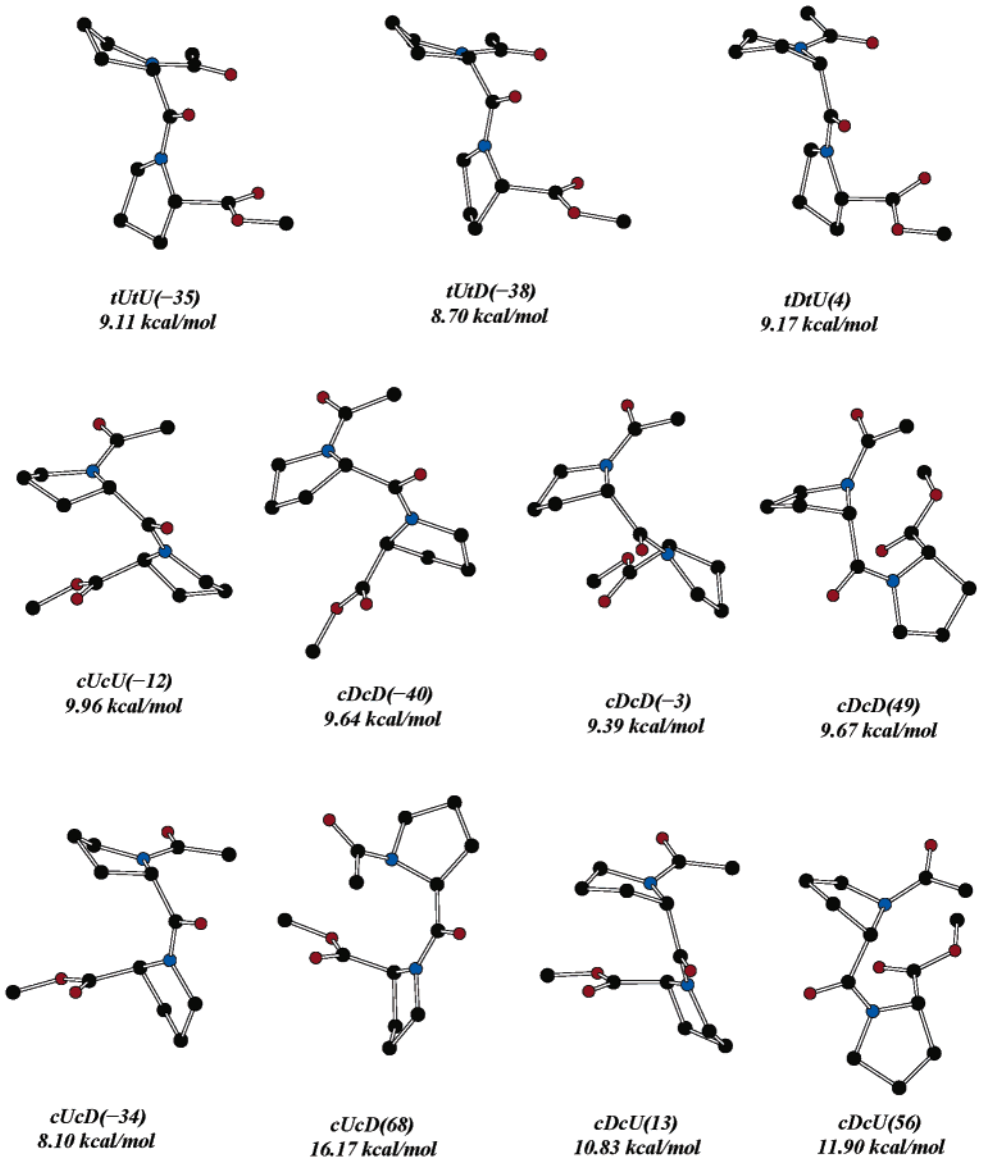


Figure 4. High-energy minima of dimers illustrating various rotamers and puckering conformations. Hydrogen atoms are not shown for clarity. Color codes for atoms: black: C; red: O; blue: N.

31G* levels. The energy differences derived from the B3LYP method were smaller than those from the RHF method. Proline hexamers composed of all *trans*-proline units (conformers 1−4 in Table 4) were found to have lower energies in the gas phase than those containing all *cis*-prolines (conformers 5−8 in Table 4). These observations are consistent with the lack of *all-cis*-proline oligomers in the PDB. The trans conformation, tDtD-hex-(148), forms an ideal left-handed 3_1 P_{II} helix (Figure 5). Conformers 2−4 in Table 4 (tDtU, tUtD, and tUtU hexamers) show similar left-handed P_{II} helices, with ϕ torsions in the range of −75° to −65° and ψ torsions in the range of 125°−152°. These variations from the ideal structure are energetically accessible and are seen in the PDB structures. The cis, endo hexamer, cDcD-hex-(165) shows an ideal, right-handed 10_3 P_I helix (Figure 6). The low-energy structures of the cDcU, cUcD, and cUcU hexamers (conformers 6−8 in Table 4) show similar right-handed P_I helices. Conformational studies of homologous β-proline oligomers have been reported, and the handedness of the β-proline oligomers is the reverse of these α-polyprolines; *trans*-α-polyprolines adopt left-handed P_{II} helices, but the *all-trans*-β-proline oligomers yield right-handed ones.[39]

The relative free energy, ΔG, for each minimum at the RHF/6-31G* level was estimated based on the calculated frequencies of the normal modes (the frequencies were scaled back by a factor of 0.89).[40] The trends in the free energy mirror those of ΔE. The ΔG values show that trans conformers are the most favorable and should be the most populated.

Here, we compare tDtD-hex-(148), our representative structure of a P_{II} helix to P_{II} structures determined by NMR (pdb code: 1JVR[41]) and X-ray crystallography (pdb file: 1F34[42]). The average ϕ and ψ torsions in the minimized trans, endo hexamer are −72.0° and 152.9°, respectively. This is in good agreement with the ϕ and ψ values from the NMR structure and the X-ray structures (Table 5). The comparison of calculated and experimental geometrical parameters for the trans, exo hexamer tUtU-hex-(133) are listed in Table

Table 4. Energies (in kcal/mol), Free Energies (in kcal/mol), and Conformational Characteristics of the Proline Hexamer Minima

conformers	characteristics[a]	ϕ (°)[b]	ψ (°)[b]	ΔE (RHF)	ΔG (RHF)[c]	ΔE (B3LYP)	handedness
1	tDtD-hex-(148)	−72.0	152.9	0	0	2.46	left
2	tDtU-hex-(135)	−74.8	125.0	2.33	1.79	0.00	left
3	tUtD-hex-(135)	−71.7	125.7	3.06	2.02	1.01	left
4	tUtU-hex-(133)	−65.7	126.1	4.31	4.98	1.00	left
5	cDcD-hex-(165)	−75.8	163.4	10.31	12.38	7.33	right
6	cUcD-hex-(161)	−68.6	163.1	14.06	15.85	10.06	right
7	cDcU-hex-(169)	−71.3	164.0	14.18	15.95	10.28	right
8	cUcU-hex-(165)	−65.8	164.5	17.88	18.96	13.00	right
9	tUtU-hex-(−35)	−66.1	−36.5	27.00	30.25	21.00	right
10	tUtD-hex-(−38)	−71.8	−26.1	29.13	31.19	23.62	right
11	tDtU-hex-(4)	−71.4	−21.6	29.47	31.64	19.68	right
12	cUcU-hex-(−12)	−70.3	−37.7	48.78	52.93	36.08	left
13	cUcD-hex-(−34)	−77.7	−30.6	49.75	53.87	37.70	left
14	cDcD-hex-(−3)	−91.0	−21.0	51.52	55.69	39.67	left
15	cDcU-hex-(13)	−80.1	−23.4	51.75	56.01	39.40	left

[a] The characteristics note the peptide rotamer, ring pucker, and ψ torsion of the minimized dimers used to build hexamers. [b] The average ϕ and ψ of the hexamers minimized with B3LYP/6-31G*. [c] The relative free energy, ΔG, was derived from the frequency calculations at the RHF/6-31G* level. The frequencies were scaled by 0.890, and the calculation was determined for $T = 298.15$ K.

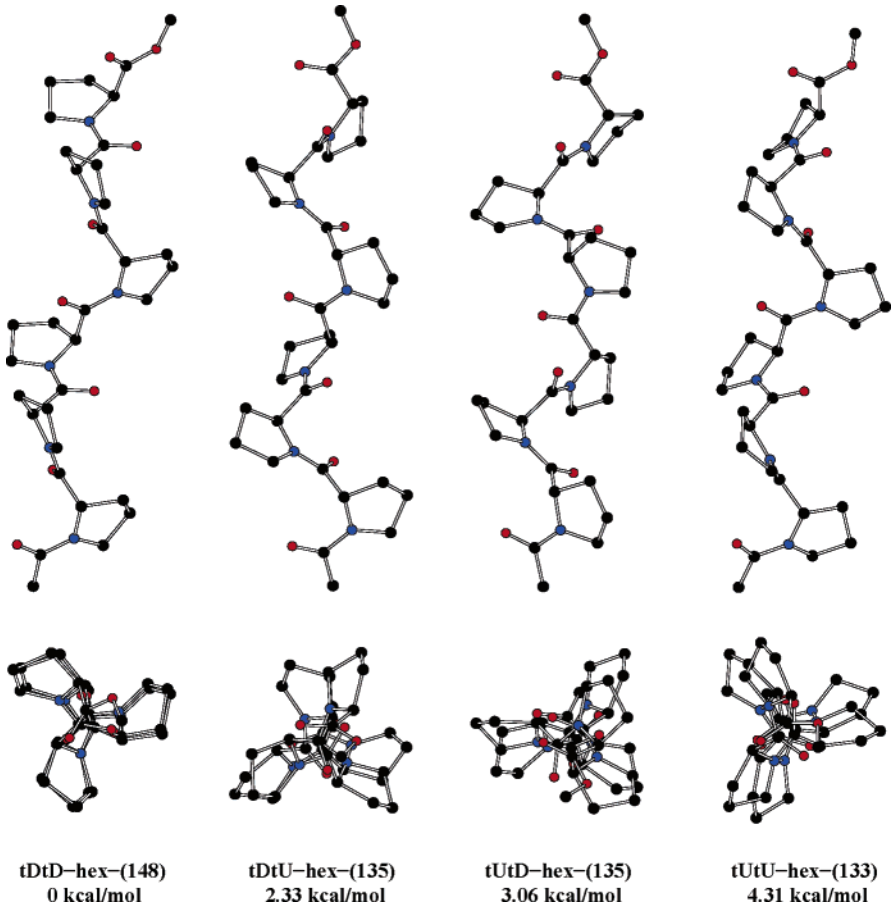


Figure 5. Side views and axial views for the low-energy, trans hexamers determined at the RHF/6-31G* level of theory. Hydrogen atoms are not shown for clarity. Color codes for atoms: black: C; red: O; blue: N.

1S. Again, the calculated average ϕ torsions are very close to those determined from the X-ray structure (pdb file: 1CF0[43]). The choice of tDtD-hex-148 and tUtU-hex-(133) was primarily based on the available NMR and/or X-ray structures with the same puckering pattern, i.e., trans, endo (tDtD-hex) and trans, exo (tUtU-hex).

Numerous experiments have shown that in polar solvents such as water, aliphatic acids, or benzyl alcohol, the cis P_I will rearrange to create the trans P_{II} form.[44,45] In contrast, the trans P_{II} will isomerize to P_I in organic solvents such as propanol or butanol. To probe this, solvation effects were estimated for the different hexamer conformations in three

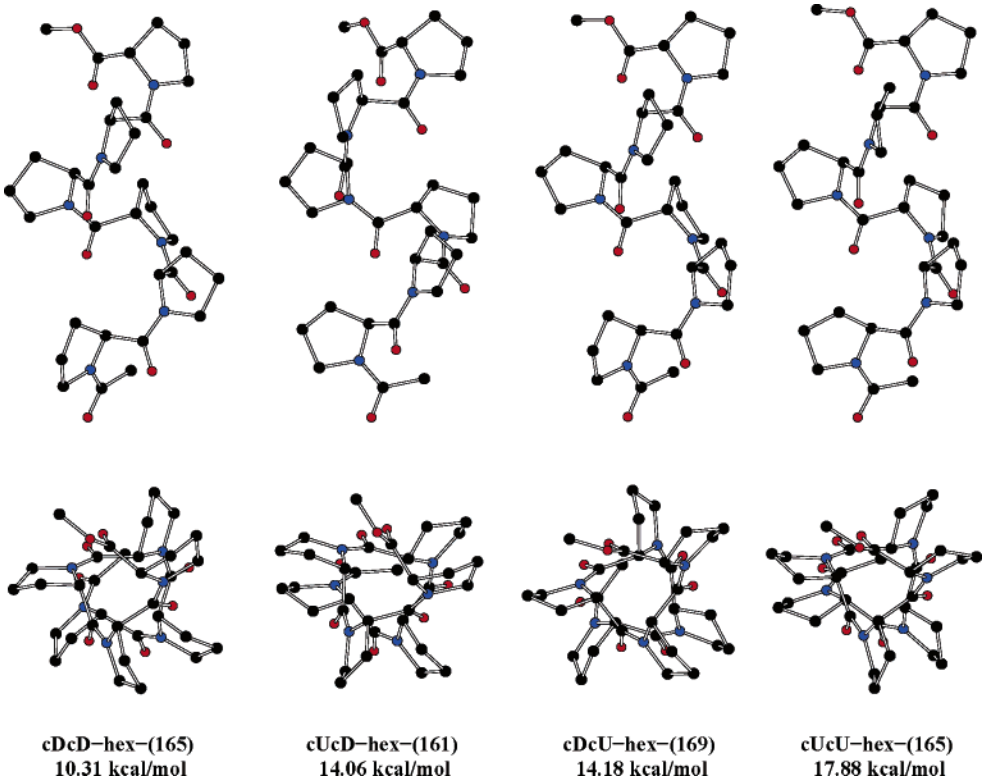


Figure 6. Side views and axial views for the low-energy, cis hexamers determined at the RHF/6-31G* level of theory. Hydrogen atoms are not shown for clarity. Color codes for atoms: black: C; red: O; blue: N.

Table 5. Comparison of Calculated Main Chain Torsion Angles (°) from a 3_1-Helical tDtD-hex-(148) with NMR and X-ray Data

	B3LYP/6-31G*			NMR (1JVR)			X-ray (1F34)		
res	ω	ϕ	ψ	ω	ϕ	ψ	ω	ϕ	ψ
1	176.7	−71.9	147.8	180.0	−75.0	159.3	179.6	−54.3	137.8
2	171.1	−73.4	153.9	180.0	−75.0	164.2	179.9	−64.4	159.4
3	171.1	−72.5	154.1	179.9	−75.0	155.0	179.7	−81.8	148.8
4	170.3	−70.5	153.7	180.0	−75.0	159.3	179.9	−61.1	147.4
5	172.1	−71.6	149.8	180.0	−75.0	171.4	179.8	−50.9	147.1
6	173.5	−72.1	158.1	180.0	−75.1	170.0			
7				179.9	−75.1	56.5			
av	172.5	−72.0	152.9	180.0	−75.0	161.8	179.8	−62.5	148.1

different environments (Table 6). The effects were calculated using the SCRF-IPCM method with increasing dielectrics for three solvents: chloroform, methanol, and water. A uniform decrease in ΔE between trans and cis low-energy conformers is observed as the solvent polarity increases. These trends can be explained by the dielectric better complementing the larger dipole moments for the cis low-energy hexamers. However, the IPCM calculations indicate that for lower-energy hexamers, the cis forms are better solvated in water and methanol than the corresponding trans forms (Table 6). This result is somewhat surprising because *trans*-proline oligomers are overwhelmingly preferred in water according to numerous experiments. However, some research groups point out that (1) the P_{II} content is solvent dependent and that the population of the P_{II} decreases in such order, water > methanol > ethanol > 2-propanol,[46] and (2) the stability of the P_{II} structure of proline oligomers is chain-length dependent. Proline oligopeptides composed of 13 Pro residues are quite stable in water, while proline hexamers and tetramers show decreasing stability due to molecular thermal fluctuation.[47]

Of course, protic solvents such as water and alcohols do not behave like simple dielectrics. There are factors other than the dipole moments that can contribute to the stability of the polyproline helices in the condensed phase. The ICPM calculations cannot account for the hydrogen bonding between the solvents and the polyproline helices. These effects strongly influence the stability of the helices. The most important point to the ICPM calculations is the fact that environmental effects overcome the large energy difference between P_I and P_{II} helices in the gas phase, making both forms stable in the condensed phase. The condensed phase also lowers the energy of conformers 9−11 which is relevant to our discussions of high-energy helices below.

The calculated IR spectrum (Table 2S in Supporting Information) from the frequency calculations show that all

Table 6. Dipole Moment (μ) and SCRF Energies (kcal/mol) of the Hexamer Minima at the RHF/6-31G* Level in the Gas Phase, $CHCl_3$, MeOH, and H_2O

conformer	characteristics[a]	μ (Debye)	ΔE (gas)	ΔE ($CHCl_3$)	ΔE (MeOH)	ΔE (H_2O)
1	tDtD-hex-(148)	4.79	0.00	0.00	0.00	0.00
2	tDtU-hex-(135)	3.04	2.33	2.70	3.28	3.25
3	tUtD-hex-(135)	3.62	3.06	2.23	2.93	2.94
4	tUtU-hex-(133)	5.62	4.31	4.97	5.58	5.49
5	cDcD-hex-(165)	25.59	10.31	2.03	−0.80	−1.32
6	cUcD-hex-(161)	26.20	14.06	5.67	3.69	3.31
7	cDcU-hex-(169)	26.31	14.18	5.73	3.17	2.86
8	cUcU-hex-(165)	26.93	17.88	8.52	5.08	4.53
9	tUtU-hex-(−35)	21.54	27.00	19.38	17.06	16.58
10	tUtD-hex-(−38)	19.32	29.13	24.49	22.98	22.68
11	tDtU-hex-(4)	19.93	29.47	22.71	20.24	19.83
12	cUcU-hex-(−12)	8.49	48.78	46.35	45.71	45.52
13	cUcD-hex-(−34)	7.86	49.75	44.93	44.04	43.48
14	cDcD-hex-(−3)	5.79	51.52	46.44	44.64	44.29
15	cDcU-hex-(13)	5.75	51.75	46.48	42.54	42.14

[a] The characteristics note the peptide rotamer, ring pucker, and ψ_1 torsion of the minimized dimers, the parents to build hexamers. The numbers in brackets denote the conformer numbers in the table.

calculated hexamers contain absorption bands at 1756 cm^{-1} which is characteristic of the ester carbonyl C═O stretch and between 1680 and 1700 cm^{-1} characteristic of the amide carbonyl stretch.[48] A strong band at 1421 cm^{-1} was also observed in all hexamers and has been reported previously for both P_I and P_{II} types of polyprolines.[49] Each of the low-energy, cis hexamers display a characteristic P_I absorption band at 960 cm^{-1} and have no P_{II} specific bands at between 670 and 400 cm^{-1}.[50] This finding suggests that the most favorable cis hexamers have the properties of P_I. Low-energy, trans hexamers have no band at 960 cm^{-1} (except tDtD-hex-148) but have bands at 400 and/or 670−675 cm^{-1}, indicating the characteristics of P_{II} helices in the most favorable trans hexamers. The higher-energy conformations from both the cis and trans hexamers have characteristic bands at 400 and 960 cm^{-1}. The calculated spectra suggest that these conformations may have some characteristics of both P_I and P_{II}.

The high-energy, trans hexamers are right-handed helices with four residues per turn (4_1), adopting a "square helix" form with a proline ring at each corner (Figure 7). We refer to this novel secondary structure for *trans*-proline oligomers as a polyproline type-III conformation (P_{III}). The square, P_{III} helix is more compact than both the P_{II} (3_1 helix) and the classic α-helix (3.6 residues per turn). P_{III} has ϕ torsions near −70° and ψ torsions of approximately −35°. This combinations of ϕ and ψ angles lie in the allowed α-helix region ($\phi \sim -57°$ and $\psi \sim -47°$), in contrast to the ϕ and ψ angles of P_{II} which are located in the β-sheet region. The conformational state for proline with ψ near −50° has been shown to be stable both experimentally[51] and computationally.[38]

It is noteworthy that trans oligomers with ψ torsions between 133° and 155° adopt a left-handed P_{II} helix, but

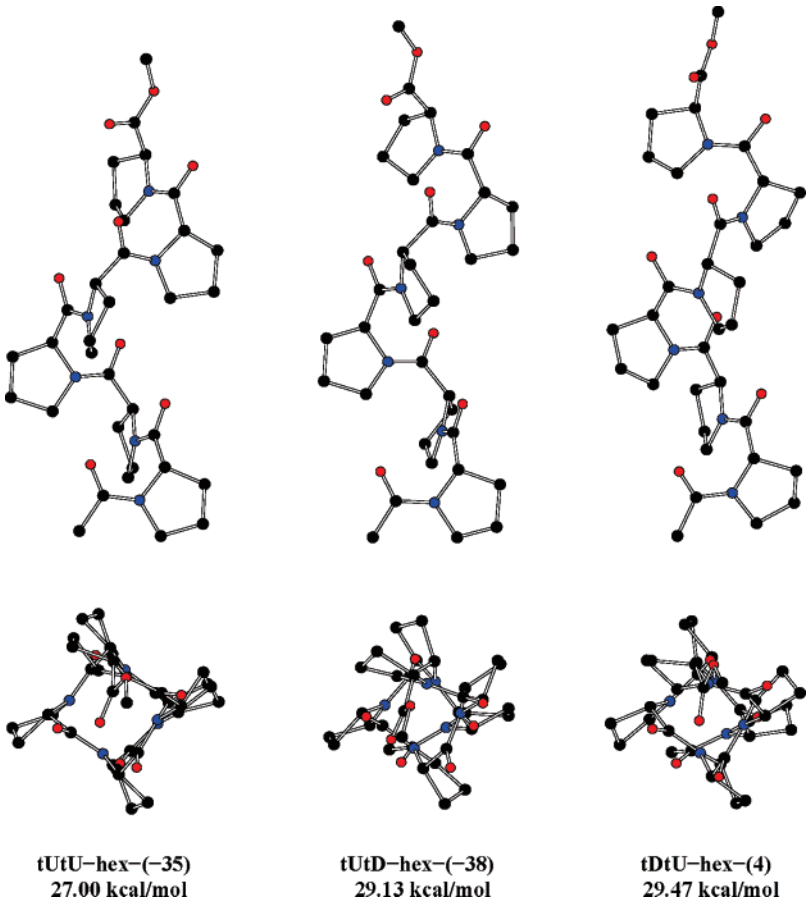


Figure 7. Side views and axial views for the higher-energy, trans hexamers determined at the RHF/6-31G* level of theory. Hydrogen atoms are not shown for clarity. Color codes for atoms: black: C; red: O; blue: N.

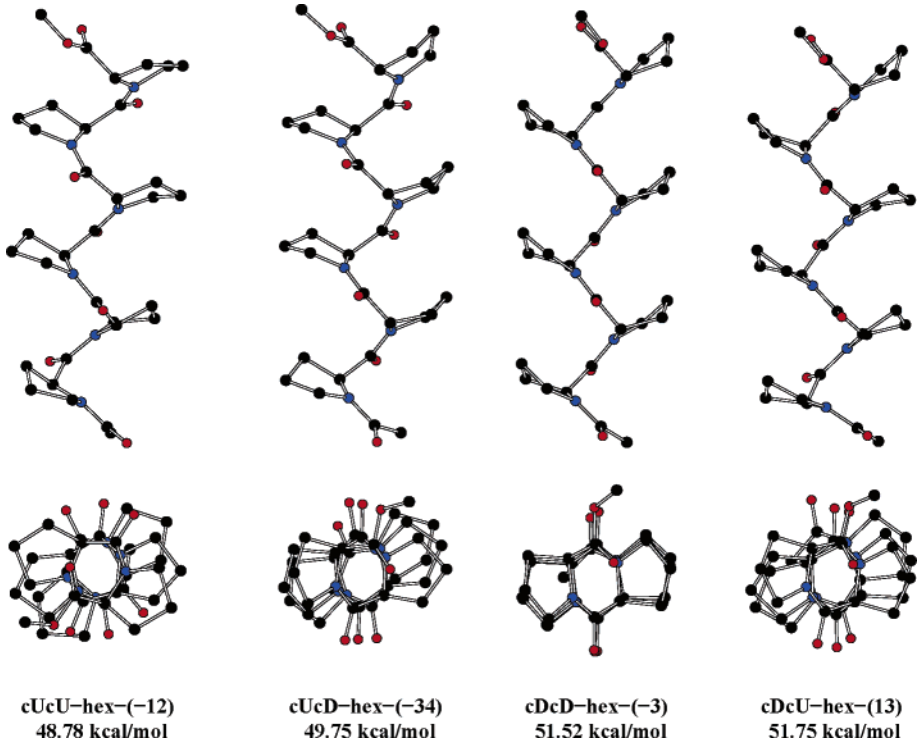


Figure 8. Side views and axial views for the higher-energy, cis hexamers determined at the RHF/6-31G* level of theory. Hydrogen atoms are not shown for clarity. Color codes for atoms: black: C; red: O; blue: N.

trans oligomers with ψ torsions near −35° give rise to the right-handed P_{III} structure. As suggested by the calculated IR spectra, the P_{III} form does indeed share characteristics of both the P_I and P_{II} forms: it has *trans*-amide rotamers similar to P_{II} and forms a right-handed helix like P_I. The handedness of the polyprolines depends not only on the peptide rotamers (cis or trans) but also on the values of the ψ torsions. We propose that the high-energy P_{III} form could exist as conformational intermediates between P_I and P_{II}. As a polyproline strand converts from the P_I to P_{II} form, it has to flip the amide rotamers and change handedness of the helix. Whether the amides flip first or the helix changes handedness first, the polyproline will have to at least partially adopt a trans, right-handed form (or a less likely cis left-handed form, see below). The ICPM calculations show that the condensed phase lowers the energies of these less favorable, trans states, making them energetically accessible and more likely than the high-energy, cis forms.

The UV-Raman spectroscopy of a 21-residue, Ala-based peptide shows a conversion from an α-helix to a P_{II} conformation.[52] Under compressive strain the polyalanine α-helix (3.66 residues per turn) was reported to transform to a π-helix (4.5 residues per turn) which is more compact than α-helix.[53] Our calculated square, P_{III} helix is very similar to the reported π-helix of polyalanine. Whether a right-handed, trans P_{III} conformation plays a role during the meltdown of the right-handed α-helix to the left-handed P_{II} conformation remains unknown. By providing the characteristics of the P_{III} conformation through calculations, it may be possible to design experiments to observe that form.

The high-energy, cis hexamers adopt conformations similar to β-strands with two residues per turn (Figure 8). They tend to adopt a slight left-handed twist if they deviate away from the ideal conformation seen for cDcD-hex-(−3). We refer to this novel secondary structure as polyproline type IV (P_{IV}). The P_{IV} sheets have ϕ torsions in the range of −90° to −70° and ψ torsions in the range of −40° to −20°, lying in the allowed α-helix region. This is different from the low-energy, cis form of P_I that has ϕ and ψ torsions in the β-sheet region of the Ramachandran plot. P_{IV} structures have relatively low dipole moments which show little stabilization in the condensed phase (Table 6). It is unlikely that this form could be observed experimentally, but it may be useful in understanding the conformational behavior of peptides or in the design of biomaterials with unique properties.

Conclusions

Our calculations provide a basis for understanding the conformational behavior of polyproline and provide an explanation for the proline oligomer distribution in the current PDB. Our calculations show that in the gas phase, *trans*-proline P_{II} helices are energetically more favorable than *cis*-polyprolines. In the condensed phases, the P_I and P_{II} forms become much closer in energy. The energy difference in ring puckering is small but slightly biased toward down-puckering. Both states would be highly populated.

To our knowledge, this is the first report of novel secondary structures for polyproline, the P_{III} and P_{IV} forms. P_{III} forms a square, right-handed helix, and P_{IV} is a β-sheet form. This is also the first report of the interconversion between left- and right-handed forms due solely to changes in the ψ torsion. Frequency calculations on the P_{III} and P_{IV} forms show that they possess the IR bands characteristic of both P_I and P_{II}. It is quite possible that the P_{III} form is an

intermediate state in the mutarotation of polyproline from P_{II} to P_I helices. Although P_{III} and P_{IV} would be less populated because of their high energy, their existence might aid in our understanding of the conformational behavior of polyproline in protein folding and provide some insight for better understanding the interconversion between P_{II} and P_I helices.

Acknowledgment. The authors are indebted to Prof. William L. Jorgensen and Dr. D. C. Lim for their generous donation of the XChemEdit program used to make many figures and to visually analyze the normal modes of optimized hexamers. The authors also thank Dr. Eugene Stewart for his critical review of this paper. This work has been supported by the National Institutes of Health (GM 65372).

Supporting Information Available: Comparison of calculated geometry parameters from a 3_1-helical tUtU-hex-(133) with X-ray data, the calculated IR frequencies for hexamers, the absolute energies (in hartrees) for proline monomers, dimers, and hexamers, the PES maps for proline monomers, and the Cartesian coordinates for all reported minima. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Zarrinpar, A.; Bhattacharyya, R. P.; Lim, W. A. *Sci. STKE* **2003**, *179*, re8., 1−10.

(2) Wiesner, S.; Stier, G.; Sattler, M.; Macias, M. J. *J. Mol. Biol.* **2002**, *324*, 807−822.

(3) Gertler, F. B.; Niebuhr, K.; Reinhard, M.; Wehland, J.; Soriano, P. *Cell* **1996**, *87*, 227−239.

(4) Freund, C.; Kühne, R.; Yang, H.; Park, S.; Reinherz, E. L.; Wagner, G. *Embo. J.* **2002**, *21*, 5985−5995.

(5) Gu, W.; Kofler, M.; Antes, I.; Freund, C.; Helms, V. *Biochemistry* **2005**, *44*, 6404−6415.

(6) Mahoney, N. M.; Janmey, P. A.; Almo, S. C. *Nat. Struct. Biol.* **1997**, *4*, 953−960.

(7) Mahoney, N. M.; Rozwarski, D. A.; Fedorov, E.; Fedorov, A. A.; Almo, S. C. *Nat. Struct. Biol.* **1999**, *6*, 666−671.

(8) Ramakrishnan, V.; Ranbhor, R.; Durani, S. *J. Am. Chem. Soc.* **2004**, *126*, 16332−16333.

(9) Whittington, S. J.; Chellgren, B. W.; Hermann, V. M.; Creamer, T. P. *Biochemistry* **2005**, *44*, 6269−6275.

(10) Hamburger, J. B.; Ferreon, J. C.; Whitten, S. T.; Hilser, V. J. *Biochemistry* **2004**, *43*, 9790−9799.

(11) Cowan, P. M.; McGavin, S. *Nature* **1955**, *176*, 501−503.

(12) Straub, W.; Shmueli, U. *Nature* **1963**, *198*, 1165−1166.

(13) Chao, Y.-Y. H.; Bersohn, R. *Biopolymers* **1978**, *17*, 2761−2767.

(14) Deber, C. M.; Bovey, F. A.; Carver, J. P.; Blout, E. R. *J. Am. Chem. Soc.* **1970**, *92*, 6191−6198.

(15) Zhang, R.; Madalengoitia, J. S. *Tetrahedron. Lett.* **1996**, *37*, 6235−6238.

(16) Counterman, A. E.; Clemmer, D. E. *J. Phys. Chem. B* **2004**, *108*, 4885−4898.

(17) Jhon, J. S.; Kang, Y. K. *J. Phys. Chem. A* **1999**, *103*, 5436−5439.

(18) Benzi, C.; Improta, R.; Scalmani, G.; Barone, V. *J. Comput. Chem.* **2002**, *23*, 341−350.

(19) Fischer, S.; Dunbrack, R. L., Jr.; Karplus, M. *J. Am. Chem. Soc.* **1994**, *116*, 11931−11937.

(20) Kang, Y. K. *J. Mol. Struct. (THEOCHEM)* **2004**, *675*, 37−45.

(21) Hudáky, I.; Baldoni, H. A.; Perczel, A. *J. Mol. Struct. (THEOCHEM)* **2002**, *582*, 233−249.

(22) Hudáky, I.; Perczel, A. *J. Mol. Struct. (THEOCHEM)* **2003**, *630*, 135−140.

(23) Czinki, E.; Császár, A. G. *Chem. Eur. J.* **2003**, *9*, 1008−1019.

(24) Ramek, M.; Kelterer, A.-M.; Nikolić, S. *Int. J. Quantum Chem.* **1997**, *65*, 1033−1045.

(25) Tanaka, S.; Scheraga, H. A. *Macromolecules* **1974**, *7*, 698−705.

(26) Bour, P.; Kubelka, J.; Keiderling, T. A. *Biopolymers* **2002**, *65*, 45−59.

(27) Vitagliano, L.; Berisio, R. A.; Mastrangelo, A.; Mazzarella, L.; Zagari, A. *Protein Sci.* **2001**, *10*, 2627−2632.

(28) Cubellis, M. V.; Caillez, F.; Blundell, T. L.; Lovell, S. C. *Proteins: Struct., Funct., Bioinformatics* **2005**, *58*, 880−892.

(29) 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 A.7*; Gaussian, Inc.: Pittsburgh, PA, 1998.

(30) Lim, D. C.; Jorgensen, W. L. In *The Encyclopedia of Computational Chemistry*; Schleyer, P. v. R., Ed.; John Wiley & Sons Ltd.: Athens, GA, 1998; Vol. 5, p 3295.

(31) Foresman, J. B.; Keith, T. A.; Wiberg, K. B.; Snoonian, J.; Frisch, M. J. *J. Phys. Chem.* **1996**, *100*, 16098−16104.

(32) Long, S. B.; Hancock, P. J.; Kral, A. M.; Hellinga, H. W.; Beese, L. S. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 12948−12953.

(33) Kang, Y. K. *J. Phys. Chem. B* **2002**, *106*, 2074−2082.

(34) Lesarri, A.; Mata, S.; Cocinero, E. J.; Blanco, S.; Lüpez, J. C.; Alonso, J. L. *Angew. Chem., Int. Ed.* **2002**, *41* (24), 4673−4676.

(35) Quan, J. M.; Wu, Y. D. J. *Theor. Comput. Chem.* **2004**, *3*(2), 225−243.

(36) Badoni, H. A.; Rodriguez, A. M.; Zamora, M. A.; Zamarbide, G. N.; Enriz, R. D.; Farkas, Ö.; Császár, P.; Torday, L. L.; Sosa, C. P.; Jàkli, I.; Perzel, A.; Papp, J. G.; Hollosi, M.; Csizmadia, I. G. *J. Mol. Struct. (THEOCHEM)* **1999**, *465*, 79−91.

(37) Kang, Y. K.; Park, H. S. *J. Mol. Struct. (THEOCHEM)* **2005**, *718*, 17−21.

(38) Mattice, W. L.; Nishikawa, K.; Ooi, T. *Macromolecules*. **1973**, *6*, 443−446.

(39) Sandvoss, L. M.; Carlson, H. A. *J. Am. Chem. Soc.* **2003**, *125*, 15855−15862.

(40) http://www.nist.gov/compchem/irikura/prog/thermo_cgi.html.

(41) Christensen, A. M.; Massiah, M. A.; Turner, B. G.; Sundquist, W. I.; Summers, M. F. *J. Mol. Biol.* **1996**, *264*, 1117−1131.

(42) Ng, K. K.; Petersen, J. F. W.; Cherney, M. M.; Garen, C.; Zalatoris, J. J.; Rao-Naik, C.; Dunn, B. M.; Martzen, M. R.; Peanasky, R. J.; James, M. N. G. *Nat. Struct. Biol.* **2000**, *7*, 653−657.

(43) Mahoney, N. M.; Rozwarski, D. A.; Fedorov, E.; Fedorov, A. A.; Almo, S. C. *Nat. Struct. Biol*. **1999**, *6*, 666−671.

(44) Steinberg, I. Z.; Berger, A.; Katchalski, E. *Biochim. Biophys. Acta* **1958**, *28*, 647.

(45) Lin, L.-N.; Brandts, J. F. *Biochemistry* **1980**, *19*, 3055−3059.

(46) Liu, Z.; Chen, K.; Ng, A.; Shi, Z.; Woody, R. W.; Kallenbach, N. R. *J. Am. Chem. Soc.* **2004**, *126*, 15141−15150.

(47) Kakinoki, S.; Hirano, Y.; Oka, M. *Polym. Bull.* **2005**, *53*, 109−115.

(48) Isemura, T.; Okabayashi, H.; Sakakibara, S. *Biopolymers* **1968**, *6*, 307−321.

(49) Johnston, N.; Krimm, S. *Biopolymers* **1971**, *10*, 2597−2605.

(50) Rabolt, J. F.; Wedding, W.; Johnson, K. W. *Biopolymers* **1975**, *14*, 1615−1622.

(51) Clark, D. S.; Dechter, J. J.; Mandelkern, L. *Macromolecules* **1979**, *12*, 626−633.

(52) Asher, S. A.; Mikhonin, A. V.; Bykov, S. *J. Am. Chem. Soc.* **2004**, *126*, 8433−8440.

(53) Ireta, J.; Neugebauer, J.; Scheffler, M.; Rojo, A.; Galván, M. *J. Am. Chem. Soc.* **2005**, *127*, 17241−17244.

CT050182T

JCTC Journal of Chemical Theory and Computation

Quantum-Chemical Design of Cryptand-like Ditopic Salt Binders

Siân T. Howard,*,† David E. Hibbs,*,‡ Angelo J. Amoroso,§ and James A. Platts§

School of Pharmacy and Medical Sciences, University of South Australia, Adelaide, South Australia 5000, Australia, Faculty of Pharmacy, University of Sydney, Camperdown, New South Wales 2006, Australia, and Department of Chemistry, Cardiff University, Cardiff CF10 3TB, Wales, U.K.

Received November 6, 2005

Abstract: Hartree−Fock, density functional, and MP2 methods are applied to the problem of designing neutral, bicyclic C_3-symmetric cages incorporating interacting anion- and cation-binding sites which strongly bind NaCl as an ion contact pair. A large number of trial ligands L and their complexes L:NaCl are tested, with the focus on maximizing binding by (i) optimizing the cavity size and shape and (ii) varying the nature of the anion- and cation-binding functionalities. The corresponding complexes L:Cl^- and L:Na^+ are also studied in some detail. An analysis of their structures and charge distributions helps to build a consistent picture of the requirements for a successful NaCl binding. The 'best' candidate ligand utilizes a tripodal triether-substituted amine $N(CH_2CH_2OR-)_3$ to bind the sodium cation; three thiourea groups in a tripodal arrangement with a 1,3,5-trisubstituted benzyl spacer group $\{C_6H_3(CH_2NHC{=}XNH-)_3$ X=O,S$\}$ to bind chloride; and a $-CH_2CH_2-$ spacer linking the two binding sites. A simple Quantitative Structure−Property analysis suggests that the binding cavity shape and size is near to the optimal one for this system.

Introduction

One of the emerging fields of interest in supramolecular chemistry is the recognition of ion pairs.[1,2] The key idea embodied in the design of appropriate multisite receptors is one of cooperativity, i.e., that the binding of one ion might facilitate either stronger or more selective binding of the other. Moreover, because a neutral host ligand incorporating an ion pair M^+X^- provides an overall uncharged system, for biochemical applications the transport properties of such neutral ditopic binders across lipophillic membranes might be superior. Sodium ions and chloride ions are the dominant cationic and anionic species in human interstitial fluid.[3] Chloride transport disfunction is associated with a number of disease states including cystic fibrosis.[4] A synthetic host which which selectively binds chloride and can pass across cell membranes might therefore have therapeutic possibilities[5] or could be used in a chloride assay. A molecule which binds NaCl selectively and reversibly could also form the basis of a *chemical* desalination/purification process for drinking water.

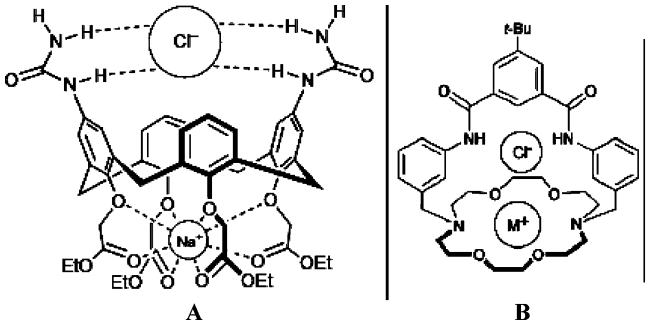
Suitable ditopic receptors might be classified into two broad categories: (i) those which bind a given pair of ions at sites remote from each other, such as the calixarene-based **A** (see Chart 1) due to Reinhoudt and co-workers[6] or (ii) ion contact pair binders such as the diamide-functionalized crown ether **B** (see Chart 1) recently reported by Smith et al.[7,8] The latter binds NaCl, KCl, and various trigonal anion combinations such as KNO_3. Moreover it has been shown to be capable of transporting the bound ion pair across a vesicle membrane.[7]

In a simplistic picture of macrocyclic ditopic receptor design (see Scheme 1) we can identify five types of elements: the anion binding sites (AB) and the anion binder cap; the cation binding sites (CB) and the cation binder cap;

* Corresponding author phone: 61883021944; fax: 61883022389; e-mail: sian.howard@unisa.edu.au (S.T.H.).
† University of South Australia.
‡ University of Sydney.
§ Cardiff University.

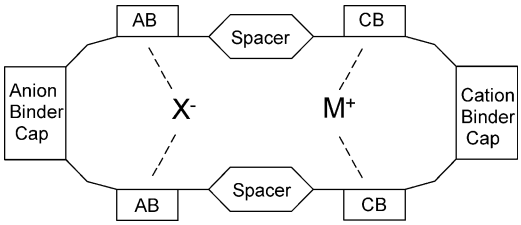
10.1021/ct050270d CCC: $33.50
Published on Web 02/02/2006

Chart 1



A B

Scheme 1

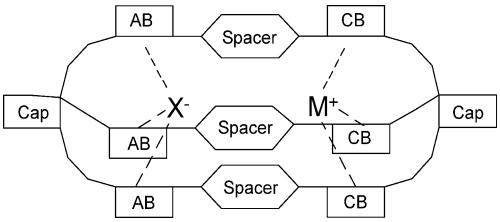


and the spacer or linker group between them. In general we might reasonably assume that a long spacer will lead to a type A system with the anion and cation bound separately and not as an ion contact pair.

In the design of new ditopic receptors, it seems clear that modeling and simulation techniques at various levels of theory have tremendous potential for quantifying the binding strength of anion and cation sites and the effect of a given spacer element on $X^-...M^+$ interactions and cooperative binding effects. Yet surprisingly, this is virgin territory in the quantum chemistry literature. Although a small number of quantum-chemical studies on anion binders have appeared in recent years,[9−16] to date just one paper has appeared on ditopic binders: the study of Geerlings et al.[17] They presented dft calculations on charged tin-containing crown ether-based host species capable of simultaneously binding Na^+ or K^+ along with SCN^- at a remote site, i.e., a type **A** ditopic binder in our simple classification scheme. The aim of this work is a computational study of *neutral* hosts which bind a salt M^+X^- as an ion contact pair, to identify potential new ditopic binders. We also spend some time considering which quantum chemical technique(s) are most appropriate for this task. This study focuses on the design of an optimal ligand for a particular ion pair (NaCl) without considering the question of selectivity, which will be the topic of a subsequent paper.

As mentioned above, at least one successful NaCl ditopic binder of type **B** has already been reported.[7,8] Although this bicyclic molecule and its complex with NaCl is small in terms of supramolecular chemistry, lacking any elements of symmetry it already represents a major challenge for high-level computational methods, where it is natural to take a 'minimalist' approach with respect to the system size and to utilize symmetry wherever possible. Initially we spent some time considering whether a small macrocyclic system such as the one shown schematically in Scheme 1 would have the desired properties. Invariably we found that a simple macrocycle either did not encapsulate the ion pair effectively,

Scheme 2



or the ions did not pair at all. This led us to focus on C_3-symmetric bicyclic cryptand-like ligands, illustrated schematically in Scheme 2. This has several advantages: (i) the extra anion-binding site significantly increases anion binding, (ii) the bicycle more effectively encloses the ion pair (bound along the C_3 axis) which both reduces the likely role any solvent molecules would play and improves the likelihood of a size-selective mechanism, and (iii) the presence of a C_3 axis makes the calculations more efficient.

Computational Methods and Procedure

The complexes were initially modeled as L:NaCl with approximate C_3 symmetry using the MM3+ force field in Macromodel.[18] Low-level quantum mechanical (HF/3-21G) geometry optimizations were then performed in C_1 symmetry, followed by harmonic frequency analyses to check for stability. Complexes which either distorted from C_3 symmetry or gave one or more imaginary frequencies were rejected at this stage. The "successful" complexes were then precisely symmetrized, and subsequent geometry optimization was performed at the HF/6-31+G* and BHandH/6-31+G* levels of theory (the latter choice of density functional is rationalized later in the Results section). Gas-phase binding energies were calculated incorporating HF/3-21G harmonic thermal energy corrections (at 298 K) with the vibrational energy component scaled by 0.89.[19] All calculations assumed singlet electronic ground states. Gaussian 03 was used for all quantum chemical calculations.[20]

To measure the binding energies of a given L there are two clear choices: (i) binding energy relative to the separated ions, which we will call BE(L:Na^+Cl^-), represented by the reaction

$$L + Na^+ + Cl^- \xrightarrow{BE(L:Na^+Cl^-)} L:\ NaCl \qquad (1)$$

and (ii)binding energy relative to molecular NaCl, which we will denote as BE(L:NaCl), represented by

$$L + NaCl \xrightarrow{BE(L:Na^+Cl^-)} L:\ NaCl \qquad (2)$$

In fact it does not much matter which one we use since they are related by a constant, which is (more or less) the energy required to form a gas-phase NaCl molecule from the separated ions: BE(L:Na^+Cl^-) = BE(L:NaCl) + D_0(NaCl).

For a given L, we can also measure its ability to separately bind Na^+ or Cl^-, i.e., the anion and cation affinities:

$$L + Na^+ \xrightarrow{BE(L:Na^+)} L:\ Na^+ \qquad (3)$$

$$L + Cl^- \xrightarrow{BE(L:Cl^-)} L:\ Cl^- \qquad (4)$$

Table 1. Ground-State Spectroscopic Data for $^{23}Na^{35}Cl$

	r_e (Å)	ω_e (cm^{-1})	D_0 (kJ/mol)[a]
MP2/6-311+G*	2.382	367	547.7
HF/6-31+G*	2.406	351	531.2
B3LYP/6-31+G*	2.391	349	545.5
BHandH/6-31+G*	2.347	371	562.5
experiment[21]	2.361	365	407.2

[a] HF/3-21G zero-point energy scaled by 0.89[18] used for all calculated values.

These binding energies are also of interest because (i) in solution all possible complexes L:NaCl(aq), L:Na^+(aq), and L:Cl^-(aq) would exist in equilibrium and (ii) knowing the magnitude of separate anion and cation affinities might help us to understand or interpret trends in the L:NaCl ion pair binding energies. Hence three gas-phase binding energies (BEs) were subsequently computed for each L, requiring several additional calculations for each compound, including the "empty" ligand L and its complexes with Na^+ and Cl^- at HF/3-21G, HF/6-31+G*, and BHandH/6-31+G* levels of theory. Note that all gas-phase binding energies reported can be trivially converted to gas-phase binding enthalpies ΔH by adding an extra RT correction term (from $P\Delta V = \Delta nRT$ with $\Delta n=1$) $\approx$ 2.5 kJ/mol at 298 K.

Results

We begin by analyzing the structure and binding of a few model systems for which MP2 calculations are feasible, to establish whether HF theory or DFT would be most appropriate for these systems.

NaCl. Modeling NaCl tells us how a particular level of theory deals with the key $X^-...M^+$ interaction. Gas-phase structural data on the $^1\Sigma^+$ ground state of NaCl are known from vibronic spectroscopy.[21] The calculations presented in Table 1 at four levels of theory shows that all of these (lower-level) methods considerably overestimate the binding energy of the molecule. BHandH and MP2 give the best bond length and vibrational frequency predictions (i.e. closest to experiment).

Complexes of *N,N′*-Dimethylurea (1a) and *N,N′*-Dimethylthiourea (1b) with Cl^-. As a model for one of the anion-binding moieties we use the symmetrically substituted molecule *N,N′*-dimethylurea **1a** and its thio-counterpart **1b** (Figure 1a,b). There appear to be no previous reports of calculations on these species (Frontera et al. have reported the MP2/6-311+G** structure and gas-phase binding energy for the urea:Cl^- complex[14,22]). There are various subtleties associated with the conformations of gas-phase urea and thiourea;[23,24] vibrational spectroscopy verifies C_2 symmetry for the lowest-energy conformations, but calculations give results which are strongly level and basis-set dependent. We find similar effects for the complexes of **1a** and **1b** and their complexes with Cl^-. **1a** is most stable in (nonplanar) C_2 symmetry at the HF/6-31+G* and BHandH/6-31+G* levels of theory. B3LYP/6-31+G* finds that **1a** is also stable C_s symmetry (unlike the other levels of theory). The urea complex **1a**:Cl^- (Table 2) is most stable in C_s symmetry for all levels of theory used with the exception of BHandH/6-31+G*, which prefers C_{2v} symmetry. The thiourea complex **1b**:Cl^- is most stable in C_{2v} symmetry for *all* levels of theory (Table 3).

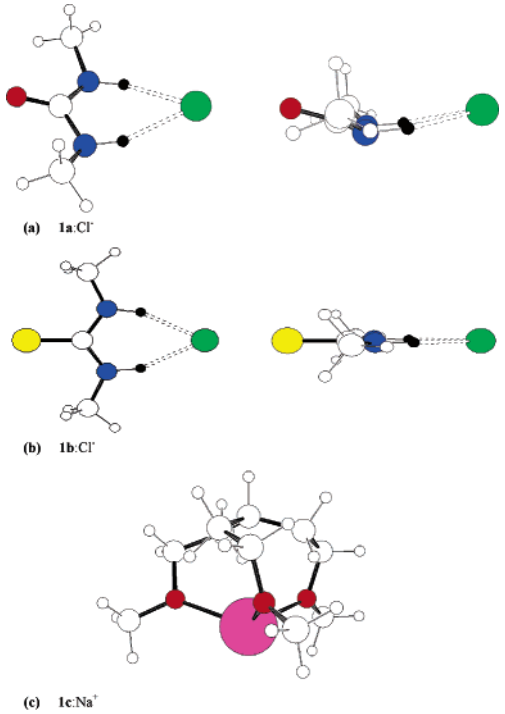


Figure 1. MP2-optimized geometries of model anion-bound and cation-bound complexes: (a) MP2/6-311+G** optimized Cl-:dimethylurea complex (C_s), two perpendicular views; (b) MP2/6-311+G** optimized Cl-:dimethylthiourea complex (C_s), two perpendicular views; (c) MP2/6-31+G* optimized H(CH_2-$CH_2OMe)_3$:Na^+ complex (C_3).

Table 2. Properties of the *N,N′*-Dimethylurea...Cl^- Complex

	r(H···Cl^-) (Å)	N−H···Cl^- (deg)	BE (kJ/mol)[a]
MP2/6-311+G** (C_s)	2.241	159.5	110.1
HF/6-31+G* (C_s)	2.486	158.9	83.8
B3LYP/6-31+G* (C_s)	2.339	158.7	96.6
BHandH/6-31+G* (C_{2v})	2.224	158.0	123.4

[a] Gas-phase binding energy including HF/3-21G thermal energies with the vibrational component scaled by 0.89.[21]

Table 3. Properties of the *N,N′*-Dimethylthiourea...Cl^- Complex

	r(H···Cl^-) (Å)	N−H···Cl^- (deg)	BE (kJ/mol)[a]
MP2/6-311+G**(C_{2v})	2.191	158.9	131.1
HF/6-31+G*(C_{2v})	2.414	160.4	109.8
B3LYP/6-31+G* (C_{2v})	2.282	159.3	119.9
BHandH/6-31+G* (C_{2v})	2.177	158.2	148.3

[a] Gas-phase binding energy including HF/3-21G thermal energies with the vibrational component sacled by 0.89.[21]

It can be seen from the data in Tables 2 and 3 that HF/6-31+G* markedly underestimates the binding energies of **1a**:Cl^- and **1b**:Cl^- and also gives much longer N−H...Cl contact distances. The B3LYP/6-31+G* and BhandH/6-31+G* binding energies are lower and higher than the MP2 result by approximately the same amount, but the BhandH

Table 4. Properties of the Complex $H(CH_2CH_2OMe)_3{:}Na^+$

	$r(O\cdots Na^+)$ (Å)	$C-O\cdots Na^+$ (deg)	BE (kJ/mol)[a]
MP2/6-31+G* (C_3)	2.339	116.5	224.5
HF/6-31+G* (C_3)	2.291	118.2	219.6
B3LYP/6-31+G* (C_3)	2.283	117.3	227.4
BHandH/6-31+G* (C_3)	2.213	115.8	262.7

[a] Gas-phase binding energy including HF/3-21G thermal energies with the vibrational component scaled by 0.89.[21]

optimized geometries are much closer to MP2 than the B3LYP results.

Complex of the Tripodal Podand $HC-(CH_2CH_2-O-Me)_3$ with Na^+. There have been a number of published studies of binding of alkali metal cations with e.g. crown ethers at the HF and MP2 levels of theory.[25] These have established that HF and MP2 give similar results for these systems, i.e., HF is a good model for cationic complexes of alkali metal ions. As a model for the common cation-binding moiety which is present in most of our ligands, we use the podand complex $HC(CH_2CH_2OMe)_3{:}Na^+$ (Figure 1c). In common with the previous studies of crown ether binding, the data in Table 2 show that HF provides a good approximation for both the geometry and binding energy of this type of complex; B3LYP is marginally poorer but still fairly close to the MP2 results. However, in this case BHandH overbinds the complex (compared to MP2) by 37 kJ/mol and correspondingly gives O...Na^+ contact distances which are some 0.12 Å too short.

In summary: although the BHandH functional seems to be the best choice for the anion...cation interaction *and* the Cl^-...urea/thiourea interaction, it is actually further from the MP2 binding energy or geometry than HF *or* B3LYP for the Na^+...ether interactions. Consequently in what follows, we present a full set of results at both HF/6-31+G* and BhandH/6-31+G* levels of theory. As a rule of thumb, we suggest that the BHandH data provide the most reliable geometries, but the "true" gas phase binding energy is bracketed by the HF and BHandH results (since these seem to consistently underestimate and overestimate binding compared to MP2, respectively).

Ditopic Salt Binders L and Their Complexes L:NaCl. A number of plausible ligands L (32 of them) were constructed, and their complexes L:NaCl were initially optimized at the HF/3-21G level. In general, two versions of each ligand have been considered, differing only by the substitution of sulfur for oxygen in the carbonyl of the amide groups of the anion-binding moiety. All of the ligands reported here contain alkyl group spacers; some of the early ligands employed benzyl spacers to make the cavity more rigid, but this tended to reduce the binding and/or distort the complex from C_3 symmetry. The structures of the ligands are summarized in Table 5; key geometrical details and binding energies are summarized in Tables 6 and 7, respectively. In the following discussion, we will focus on the DFT (BHandH) results (we note that the HF/6-31+G* equivalent to Table 6 is supplied as Supporting Information, Table 6S). The "raw" electronic and thermal energy data on which the quantities in Tables 1−7 are based are also provided as Supporting Information (Tables 1S−5S). The BhandH/6-31+G* geometry-optimized coordinates of all species are also provided in the Supporting Information (Tables 7S−62S).

The first pair of complexes to be stable and show the desired properties were **1**:NaCl and **2**:NaCl (Figure 2a,b), which employ a single −CONH− or −CSNH− amide/thioamide anion binding group and $-CH_2-CH_2-CH_2-$ spacer in each arm of the ligand. NaCl is clearly bound as a contact ion pair with a bond length $r(Na-Cl) \approx 2.37$ Å, almost identical to the gas-phase NaCl molecule at the same level of theory (Table 1). The NaCl binding energy of this ligand is 80 kJ/mol for the triamide **1** and 116 kJ/mol for the trithioamide **2**. Both HF and DFT levels of theory predict stronger binding for the thioamide ligand, and this turns out to be a consistent feature for all the successful pairs of amide/thioamide ligands. The H-bonding geometry of this ligand seems to be particularly favorable, with slightly shorter and more linear N−H...Cl contacts in **2**:NaCl, consistent with its higher binding energy. However, the acute $C-O-Na^+$ angles (83°−84°) at the cation-binding pocket of both ligands suggest a strained geometry this part of the ligand (this is 30° lower than the equivalent angles in the podand complex,

Table 5. Composition of the Various Ditopic Binders

	anion binder cap	anion-binding moiety	spacer	cation-binding moiety	cation binder cap
1	HCR_3	$-CH_2CONH-$	$-CH_2CH_2CH_2-$	$-OCH_2CH_2-$	R_3CH
2	HCR_3	$-CH_2CSNH-$	$-CH_2CH_2CH_2-$	$-OCH_2CH_2-$	R_3CH
3	HCR_3	$-NHCONH-$	$-CH_2CH_2CH_2-$	$-OCH_2CH_2-$	R_3CH
4	HCR_3	$-CH_2NHCONH-$	$-CH_2CH_2-$	$-OCH_2CH_2-$	R_3CH
5	BzR_3	$-CH_2NHCONH-$	$-CH_2CH_2-$	$-OCH_2CH_2-$	R_3CH
6	BzR_3	$-CH_2NHCSNH-$	$-CH_2CH_2-$	$-OCH_2CH_2-$	R_3CH
7	BzR_3	$-CH_2CH_2NHCSNH-$	$-CH_2CH_2-$	$-OCH_2CH_2-$	R_3CH
8	BzR_3	$-CH_2CH_2NHCONH-$	$-CH_2CH_2-$	$-OCH_2CH_2-$	R_3CH
9		$H_3C-NHCONH-$	$-CH_2CH_2-$	$-OCH_2CH_2-$	R_3CH
10		$H_3C-NHCSNH-$	$-CH_2CH_2-$	$-OCH_2CH_2-$	R_3CH
11	TACN[a]	$-CH_2NHCSNH-$	$-CH_2CH_2-$	$-OCH_2CH_2-$	R_3CH
12	TACN[a]	$-CH_2NHCONH-$	$-CH_2CH_2-$	$-OCH_2CH_2-$	R_3CH
13	BzR_3	$-CH_2NHCSNH-$	$-CH_2CH_2-$	$-OCH_2CH_2-$	R_3N
14	BzR_3	$-CH_2NHCONH-$	$-CH_2CH_2-$	$-OCH_2CH_2-$	R_3N

[a] 1,4,9-Triazacyclononane.

Table 6. BHandH/6-31+G* Geometry-Optimized Data for the Complexes L:NaCl

	r(H···Cl$^-$) (Å)	N−H···Cl$^-$ (deg)	r(O···Na$^+$) (Å)	r(Na···Cl) (Å)	C−O···Na$^+$ (deg)	q(Na$^+$) (au)[a]	q(Cl$^-$) (au)[a]
1	2.177	143.1	2.255	2.368	82.8	+0.315	−0.959
2	2.124	146.2	2.239	2.374	83.8	+0.399	−1.007
3	3.228, 3.363	83.3, 73.1	2.211	2.383	91.2	+0.473	−0.686
4	2.591, 3.018	97.8, 88.1	2.275	2.325	82.4	+0.310	−1.076
5	2.465, 2.429	141.2, 136.0	2.155	2.393	107.6	+0.299	−0.605
6	2.461, 2.242	144.1, 151.2	2.152	2.401	104.6	+0.392	−0.616
7	2.445, 2.285	150.3, 159.0	2.166	2.453	112.8	+0.569	−0.643
8	2.401, 2.364	152.2, 157.3	2.179	2.476	112.8	+0.333	−0.588
9	2.367, 2.376	146.7, 147.8	2.226	3.448	124.2	+1.588	−0.681
10	2.314, 2.353	148.6, 150.6	2.237	3.134	121.7	+0.214	−0.642
11	2.533, 2.194	129.0, 142.0	2.156	2.391	107.6	+0.451	−0.697
12	2.525, 2.202	132.1, 145.7	2.147	2.384	107.8	+0.400	−0.671
13	2.235, 2.319	162.1, 157.5	2.279	2.527	106.4	+0.462	−0.787
14	2.253, 2.403	152.6, 161.8	2.276	2.523	107.5	+0.332	−0.726

[a] Mulliken charges.

Table 7. Binding Energies of Various Complexes at Different Levels of Theory (kJ/mol)

	HF/3-21G[a]			HF/6-31+G*[a]			BHandH/6-31+G*[a]		
	L:NaCl	L:Na$^+$	L:Cl$^-$	L:NaCl	L:Na$^+$	L:Cl$^-$	L:NaCl	L:Na$^+$	L:Cl$^-$
1	121.9	370.7	118.5	−18.8	171.9	91.0	77.8	228.9	180.8
2	185.1	326.7	211.5	13.6	135.2	147.2	116.0	195.4	236.7
3	135.8	414.9	45.3	−31.9	205.1	6.6	66.1	253.5	92.3
4	53.3	480.6	13.1	−83.9	308.0	30.8	26.8	375.9	116.4
5	256.1	565.2	204.0	98.4	362.2	171.3	199.8	418.5	264.4
6	281.9	444.9	274.8	114.2	308.6	207.9	214.7	364.9	294.8
7	330.0	463.9	260.0	172.0	309.0	191.1	271.1	390.3	293.5
8	314.0	397.0	212.2	165.0	255.8	171.0	269.9	311.2	272.6
9	387.6	435.2	199.3	208.5	294.6	187.5	285.5	328.3	218.0
10	404.3	466.0	291.6	214.2	297.0	229.2	338.4	407.2	296.9
11	257.8	433.6	216.6	83.2	238.9	154.8	185.6	361.9	247.5
12	189.4	420.0	115.7	70.5	259.2	125.7	126.1	309.8	176.1
13	365.9	494.2	232.5	206.0	326.5	165.0	320.6	456.5	277.5
14	368.9	503.4	207.1	215.7	311.9	172.0	327.4	416.4	268.2

[a] Gas-phase binding energy including HF/3-21G thermal energies with the vibrational component scaled by 0.89.[21]

see Table 3). This suggests that the ligand cavity is too small; higher NaCl binding energies could be achieved if the sodium ion can relax further toward the center of the cavity.

It was reasoned that switching the anion-binding moiety to a urea/thiourea fragment ought to increase binding at the anion-binding end of L. A simultaneous reduction of the length of the spacer $-CH_2-CH_2-CH_2-$ should also have the effect of "pulling" the sodium ion toward the cavity center. These ideas were first tested in complexes **3**:NaCl and **4**:NaCl (Figure 2c,d). In **3**:NaCl which retains the longer (propyl) spacer, the urea groups are evidently too distant from the chloride anion to bind. Instead the Cl$^-$ ion remains in close contact with Na$^+$, r(Na−Cl) ≈ 2.38 Å. However the strain at the cation binding pocket is partly removed, as the C−O−Na$^+$ angles increases to 92°. In **4**:NaCl, removing one methylene group from each spacer and reinserting it between the urea groups and the R_3CH cap does facilitate N−H...Cl$^-$ hydrogen bond formation for the half of the urea fragments; but the other remaining urea N−H bonds are still too far from Cl$^-$ to engage in hydrogen bonding, and the C−O−Na$^+$ angles are once again acute at ≈83°. The NaCl binding energies of 66 and 27 kJ/mol for **3**:NaCl and **4**:NaCl, respectively, are very low, and the fact that **3** binds more strongly than **4** despite the presence of weak N−H...Cl hydrogen bonds in the latter suggests that the partial removal of strain in the cation pocket of **3**:NaCl was energetically a more important consideration.

One of the striking features of complexes **3**:NaCl and **4**:NaCl is that the N−H bonds of the urea moieties point away from the center of the cavity. This led us to conclude that the $HC(CH_2)_3$ anion binder cap is too small to permit effective urea...Cl$^-$ binding. Several possible alternative cap functional groups with potential C_3 symmetry were explored at this point, including cyclohexyl and triazacyclohexyl groups, but the HF/3-21G optimized structures were strongly distorted from C_3 and furthermore did not improve the anion binding. The benzyl and triazacyclononyl groups were much more successful, and these are discussed below.

In complexes **5**:NaCl and **6**:NaCl (Figure 2e,f) the urea/thiourea anion-binding moieties are bonded to a 1,3,5-trisubstituted benzyl spacer, $C_6H_3(CH_2)_3$. It is immediately apparent that this has the effect of opening up the anion binding end of the cavity to the extent that six N−H...Cl$^-$ hydrogen bonds are formed in both amide and thioamide versions of the ligand. The geometry at the Na$^+$ cation has also become much more favorable for optimal binding, with C−O−Na$^+$ angles of 108° (105° in the trithiourea ligand). Despite the presumably much stronger binding of Cl$^-$, the NaCl moiety remains as an ion contact pair with bond length r(Na−Cl) ≈ 2.4 Å. The overall NaCl binding energies of

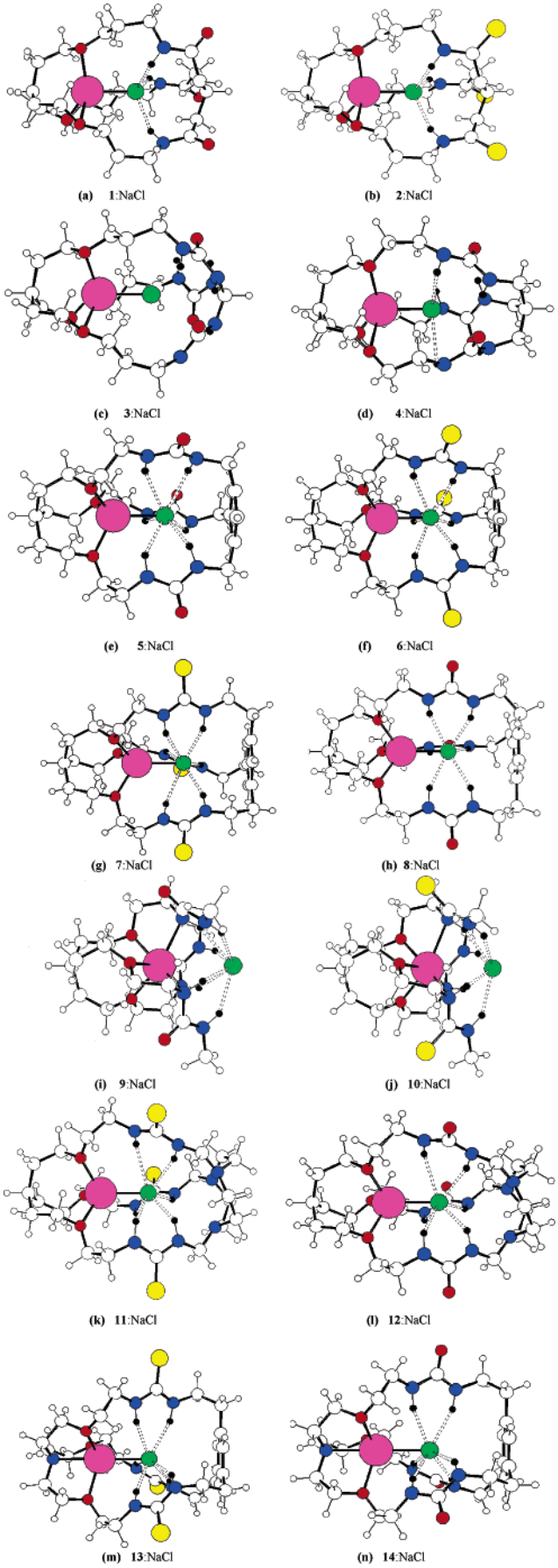


Figure 2. BHandH/6-31+G* geometry-optimized structures of the **L**:NaCl complexes (**L** = hosts **1**−**14** as described in the main text).

200 kJ/mol (215 kJ/mol for the trithiourea ligand) show a substantial improvement over the previous candidate ligands. Thus it appears that this is already an excellent candidate for a ditopic NaCl binding ligand. We note however that the hydrogen bonding geometry at Cl^- could still bear some optimization: the H...Cl^- distances in the triurea ligand are quite long at ≈2.45 Å (much more asymmetric in the trithiourea with 2.46 Å and 2.24 Å), and the N−H...Cl^- angles are still generally below 150°. Careful inspection of Figure 2e,f also shows that the benzyl caps appear to be in slightly strained (nonplanar) geometries which probably decrease the effective NaCl binding energies of the complexes.

In **7**:NaCl and **8**:NaCl the effect of further expanding the cavity at the anion-binding pocket has been explored by including an extra methylene spacer in the anion binding cap, which is now $C_6H_3(CH_2CH_2)_3$. The details of the hydrogen bonding geometries in Table 6 for both the urea and thiourea complexes show that this has been effective: the H-bonds are shorter in the trisurea case (although not much different to **6**:NaCl for the tristhiourea complex **7**:NaCl); but all N−H...Cl^- angles are now above 150° (i.e. more linear) in both complexes. The benzyl fragments are now quite planar. The C−O−Na^+ angles of ≈113° are very close to the value of ≈116° found for the podand complex $H(CH_2CH_2OMe)_3$:Na^+, which suggests that ligands **7** and **8** have a a close-to-ideal geometry for the Na^+ binding pocket. A substantial increase in overall NaCl binding energy is again seen, and this time the urea and thiourea complexes give almost identical binding (≈270 kJ/mol).

The complexes **9**:NaCl and **10**:NaCl represent our only departures from a closed bicyclic cage design—here we have explored the effect of an open-ended anion-binding moiety, i.e., a podand instead of a cryptand. Examining the geometry-optimized structures (Figure 2i,j) it is immediately evident that these complexes are qualitatively different to all the others: the r(Na−Cl) separation has stretched well beyond the gas-phase value of Å, reaching 3.45 Å in the carbonyl complex. Another feature unique to this pair of complexes is that the sodium cation is η^6-coordinated, with the nitrogens of adjacent urea/thiourea groups acting as electron donors. This fundamental change in cation coordination results in Mulliken charges which are quite out of line with the 0.3−0.6 range seen in all other complexes: an extremely high $q(Na^+) = +1.59$ for **9**:NaCl and paradoxically the lowest value seen, $q(Na^+) = +0.24$ for **10**:NaCl. In short, the effect of opening the cavity at the anion-binding end has produced a ditopic binder in which the Na^+ and Cl^- ions are no longer a contact pair, due to the moreover, this effect is accompanied by a substantial increases in overall binding energy (288 and 338 kJ/mol). We note however that the anion in these open-ended complexes would not be effectively shielded from solvent molecules, so it is questionable whether this exceptionally high binding energy would actually be achieved in the aqueous phase.

In **11**:NaCl and **12**:NaCl the effect of replacing the trisubstituted benzyl spacer with a N,N′,N″-functionalized triazacyclononane is explored. This is a common ligand in coordination chemistry.[26] We consider it here on the basis of its potential C_3 symmetry as a design element in these

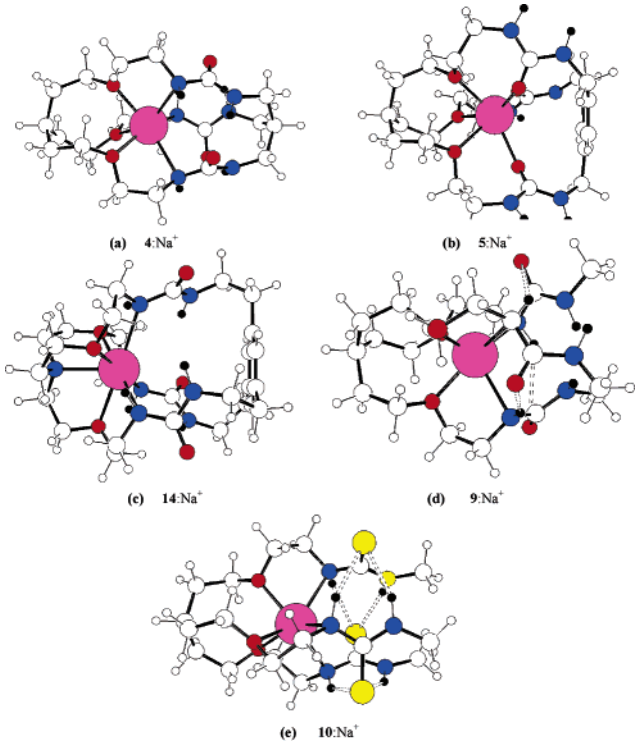


Figure 3. BHandH/6-31+G* geometry-optimized structures of selected complexes **L**:Na^+, illustrating the different modes of coordination obtained.

ligands, and because of its well-known utility in supramolecular design. It was found that, although the complexes **11**:NaCl and **12**:NaCl were stable in C_3, the free ligands distorted considerably to C_1. (For the 14 ligands reported in this paper, **11** and **12** are the only cases where the *free* ligands are not C_3-symmetric.) In the complexes with NaCl, the introduction of the TACN functionality has an adverse effect on the hydrogen bonding geometry, with less linear and generally longer hydrogen bonds than those of the previously discussed complexes. The much lower binding energies of 186 kJ/mol (trithioamide) and 126 kJ/mol (triamide) presumably reflect this poorer H-bonding geometry.

In **13**:NaCl and **14**:NaCl the effect of adding an additional coordination site for the sodium cation is explored, by simply replacing the axial CH of the cation binding caps of **7** and **8** with a nitrogen heteroatom. It might also be argued that such ligands are easier to synthesize than some of the preceding ones, using e.g. triethanolamine as a starting material. The optimized structures (Figure 2m,n) demonstrate that the sodium ion is indeed η^4-coordinated, and the complexes retain C_3 symmetry. Although the Na...Cl contact distances are slightly longer in these complexes, r(Na−Cl) $\approx$ 2.53 Å, the anion binding is not at all adversely affected; in fact the N−H...Cl− H-bonds are some of the shortest and are closer to being linear than any of the other the complexes reported here. The NaCl binding energies for this pair of ligands are very similar and are the largest obtained for any of the ligands *except* for the podand complex **10**:NaCl.

The Complexes L:Na^+. The η^6-coordination Na^+ mode which was found only in the two podand ditopic complexes **9**:NaCl and **10**:NaCl is in fact seen in a number of the complexes of these ligands with just a sodium cation (**4**:Na^+, **5**:Na^+, **7**:Na^+, **9**:Na^+, **10**:Na^+, **11**:Na^+, and **12**:Na^+). Two representative examples **4**:Na^+ and **5**:Na^+ are shown in Figure 3a,b. Figure 3a depicts the first type of coordination, involving the three ether oxygens and three nitrogen atoms of the urea groups; this occurs in **4**:Na^+, **9**:Na^+, **10**:Na^+, **11**:Na^+, and **12**:Na^+. Figure 3b depicts a second type of coordination seen in two cases (**5**:Na^+ and **7**:Na^+) where carbonyl(thiocarbonyl) oxygens(sulfurs) are acting as additional donors. Where it occurs, the switch from η^3- to η^6-coordination is unsurprisingly accompanied by an increase in L:Na^+ binding energy (for example, the 376 kJ/mol binding energy of **4**:Na^+ is much higher than the 254 kJ/mol value for η^3-coordinated **3**:Na^+), and this explains most trends seen in Table 5. Complexes **13**:Na^+ and **14**:Na^+ actually have η^7-coordination at Na^+ due to the extra capping nitrogen heteroatom (see the example in Figure 3c). Another feature which also enhances the stability of several complexes (**4**:Na^+, **9**:Na^+, **10**:Na^+, **12**:Na^+, and **13**:Na^+) is the formation of intramolecular N−H...O or N−H...S hydrogen bonds: the example (two examples are shown in Figure 3d,e). We note that, in general, the best ditopic NaCl binders also happen to be the best binders for Na^+ alone, based on these gas-phase binding energy considerations.

The Complexes L:Cl^-. The full H-bonding geometrical data for these complexes are reported in Table 7S. Although **1**:Cl^- and **2**:Cl^- only form three N−H...Cl H-bonds, these bonds are quite linear and short (see Figure 3a), hence the chloride binding energies of 181 and 236 kJ/mol (the latter for the thiourea species) are fairly high. The complex **3**:Cl^- is the 'odd one out' because N−H...Cl hydrogen bonds do not form at all due to competing N−H...O(carbonyl) hydrogen bonding interactions (see Figure 3b), so unsurprisingly its chloride binding energy is the lowest. Of the remaining complexes, **5**:Cl^-, **6**:Cl^-, **7**:Cl^-, **8**:Cl^-, **11**:Cl^-, **12**:Cl^-, **13**:Cl^-, and **14**:Cl^- are all fairly similar in that they contain three pairs of N−H...Cl H-bonds with geometries ranging from N−H...Cl = 124° − 165° and r(H...Cl) ranging from 2.21 Å to 2.51 Å. So although the "ideal" H-bonding distance of 2.17 Å (from the *N,N*-dimethylthiourea:Cl^- complex) is never quite attained for these various bicyclic ligands, several have N−H...Cl^- angles higher than the 158° obtained for *N,N*-dimethylthiourea:Cl^-. The thiourea versions of these ligands have consistently higher chloride binding energies than their urea-based equivalents, the strongest complex being found for **6**:Cl^- (shown in Figure 3c). The two remaining (podand) complexes **9**:Cl^- and **10**:Cl^- also bind chloride strongly (in fact **10**:Cl^- has an almost identical binding energy to **6**:Cl^-, see Figure 4), but both have strongly distorted to C_1 symmetry (see Figure 3d).

Discussion − Quantitative Structure−Property Relationships

The parameters listed in Table 6 have been used to construct various quantitative structure−property relationships using least-squares models of the BHandH/6-31+G* binding energies for the complexes L:NaCl. After some experimentation we found that the most efficient linear model of the binding energy is a three-parameter model involving one distance r(O···Na^+), which we will denote as d, and two

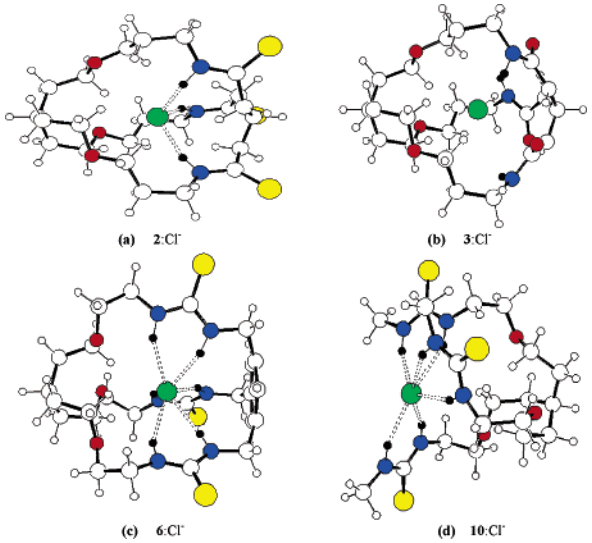


Figure 4. BHandH/6-31+G* geometry-optimized structures of selected complexes **L**:Cl⁻, illustrating the different modes of coordination obtained.

angles: N−H···Cl⁻ and C−O...Na⁺, which will be denoted as θ and ω, respectively

$$\mathrm{BE}(\theta, d, \omega) \approx -2015.1 + 1.636\,\theta + 632.41\,d + 5.699\,\omega \text{ kJ/mol} \quad (5)$$

where the distance d is in Å and the angles are in degrees. This model delivers an R^2 value of 0.91 and an rms error on the predicted binding energy of 36.2 kJ/mol for all 14 data. The remaining parameters, including the r(Na···Cl) separation and the partial (Mulliken) charges $\{q(\mathrm{Na}^+), q(\mathrm{Cl}^-)\}$ do not appear to play an important role in determining the binding energy and are also strongly linearly correlated with the three parameters in (5).

Although eq 5 provides a rough estimate of the ion pair binding energy for a given structure, the coefficients in this linear model have no physical significance. We hypothesize that there is some *optimal* cavity size and shape which will lead to the highest possible ion pair binding energy. This must be strictly true if **1**−**14** all contained an identical arrangement of binding heteroatoms but differed only in the size and shape of the binding pocket. This is not strictly the case for our molecules, for several reasons: (i) some molecules contain ureas while others contain thioureas; (ii) **1** and **2** have amide/thioamide groups instead of ureas or thioureas; and (iii) **11** and **12** have an additional capping nitrogen atom which also plays a role in cation binding. We can nevertheless explore this hypothesis by attempting to fit the simplest nonlinear model to the ion pair binding energies which is consistent with the notion of an optimal cavity size/shape. So we choose the same three structural parameters which were indicated to be most important from the linear model and rewrite the binding energy in quadratic form

$$\mathrm{BE}(\theta, d, \omega) \approx \mathrm{BE}_{\max} - k_1(\theta - \theta_0)^2 - k_2\,(d - d_0)^2 - k_3\,(\omega - \omega_0)^2 \quad (6)$$

where $\mathrm{BE}_{\max}$ represents the maximum achievable binding energy; $\{k_1, k_2, k_3\}$ are pseudoforce constants; and $\{\theta_0, d_0, \omega_0\}$ represent "optimal" values of these three structural parameters. Although there are seven parameters in this model, using Mathematica5[28] we find that it is possible to obtain a stable fit of eq 6 to our 14 ion pair binding energies:

$$\mathrm{BE} \approx 353.24 - 0.011\,(\theta - 179.6)^2 - 9553.5\,(d - 2.268)^2 - 0.219\,(\omega - 116.1)^2 \text{ kJ/mol} \quad (7)$$

Equation 7 approximates the exact BhandH/6-31+G* ion pair binding energies with an rms error of 21 kJ/mol; moreover, the fitted parameters have a direct physical interpretation. The suggested optimal values are revealing. The r(O...Na⁺) distance of ≈2.27 Å is attained in two of our complexes (**13**:NaCl and **14**:NaCl), but in most of them it is significantly shorter. The optimal C−O...Na⁺ angle is almost identical to the value obtained for $HC(CH_2CH_2OMe)_3$:Na⁺ at the same level of theory; this angle is several degrees lower in all but two of our complexes (**9**:NaCl and **10**:NaCl). Finally, the optimal N−H...Cl⁻ angle is suggested to be as close to 180° as possible. These observations taken together suggest that it would be beneficial to further increase the size of the NaCl binding pocket. However, because the fitted value of $\mathrm{BE}_{\max}$ is only 15 kJ/mol higher than our best value (for **10**:NaCl), the increase in binding energy following this modification would probably be limited.

We also considered linear models of the L:Cl⁻ binding energy. (We did not attempt to model the L:Na⁺ binding energy because of the considerable variation in binding modes seen across the 14 compounds). For the L:Cl⁻ complexes, linear models do not appear to be very useful, the best (two-parameter) model involving r(H...Cl) and N−H...Cl delivers only $R^2 = 0.797$ and an rms error of 32.4 kJ/mol for all 14 data.

Conclusions

The preliminary studies of model anion and cation binding systems showed that ditopic binders are challenging systems for quantum chemical calculations. Density functional methods are currently the only way forward for systems of this size, if electron correlation effects are to be taken into account, but choosing a density functional which accounts for both anion-binding and cation-binding interactions with similar accuracy is problematic. Nevertheless, we believe that this study has led to a number of important conclusions regarding the design of potential ditopic salt binding ligands.

(i) It certainly is possible to design cryptand-like, C_3-symmetric cages with adjacent anion and cation binding sites leading to very strong ion pair binding (binding energies in excess of 300 kJ/mol relative to free NaCl). A systematic design process using quantum chemical methods has provided several excellent candidate host molecules. An ethyl spacer between anion- and cation-binding pockets was found to be optimal for maximizing the anion−cation interaction. The introduction of a *rigid* anion binder cap in the form of a 1,3,5-trisubstituted benzene proved to be essential in providing the correct spacing for the ureas/thioureas to bind effectively to the chloride.

(ii) Thioamides and thiourea are generally better ligands for Cl^- binding than the equivalent amide and urea ligands, which up to this point have been preferred synthetically. This is probably linked to our observation that Cl^-:dimethylthiourea complexes seem to prefer a more coplanar geometry than the analogous Cl^-:dimethylurea complexes (i.e. this facilitates stronger H-bonds).

(iii) A nitrogen heteroatom for the cation-binding pocket also plays a role in cation binding. An equivalent neutral, C_3-symmetric capping moiety for the anion-binding end of the host molecule might be *s*-triazine, which according to Frontera et al. has significant $Cl^-...\pi$ interactions.[13,21] We have not attempted any calculations using this fragment as an anion-binding cap because of the uncertainty associated with how well dft methods can model the relatively exotic halide anion...π interactions.

(iv) The size and shape of the cavity in which the anion and cation sit is also crucial for optimizing the NaCl binding energy. Although selectivity for NaCl over other ion pairs has not been explicitly considered here, it seems likely that our best bicyclic host **14** would be selective for NaCl because of the considerable time spent tuning the cavity size for this particular pair of ions during the design process.

(v) Trends in binding energies of the associated $L:Na^+$ and $L:Cl^-$ complexes are complex because coordination modes of the ions are not consistent across the series of compounds. It was originally anticipated that the difference between BE(L:NaCl) and the sum of BE($L:Na^+$) and BE($L:Cl^-$) could be used to measure of 'cooperativity effects' for the binding of the ion-pair: but this is not effective because of these complicated trends in the series BE($L:Na^+$) and BE($L:Cl^-$), due to varying coordination modes.

We should finally add a word on the synthetic feasibility of the ditopic hosts we have designed. The nitrogen-capped polyether podand form of our optimal system is already familiar from the very well-developed field of cryptand chemistry[27] (and we note that suitable starting products such as triethanolamine are off-the-shelf chemicals). At the anion-binding end of the system, the podand based on a 1,3,5-trithiourea substituted benzene is not (to our knowledge) a known ligand, but functionalizing benzyl groups with ureas is fairly straightforward and has been utilized in known ditopic binders such as the one mentioned earlier by Reinhoudt and co-workers.[5] The "strapping" of these two elements to make the bicyclic system may not be at all trivial but could again draw on the very extensive knowledge base on this topic that has already been accumulated by researchers in the cryptand field.

Acknowledgment. S.T.H. and D.E.H. thank the Australian Research Council for funding (Grant DP0556144).

Supporting Information Available: HF/6-31+G* equivalent to Table 6 (Table 6S), the "raw" electronic and thermal energy data on which the quantities in Tables 1−7 are based (Tables 1S−5S), and BhandH/6-31+G* geometry-optimized coordinates of all species (Tables 7S−62S). This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Kirkovits, G. J.; Shriver, J. A.; 1, Gale, P. A.; Sessler, J. L. *J. Incl. Phenom. Mac. Chem.* **2001**, *41*, 69−75.

(2) Gale, P. A. *Coord. Chem. Rev.* **2003**, *240*, 191−221.

(3) Bretag, A. *Life Sci.* **1969**, *8*, 319−329.

(4) Rosenstein, B. J.; Zeitlin, P. L. *Lancet* **1998**, *351*, 277−282.

(5) (a) Gao, L.; Broughman, J. R.; Iwamoto, T.; Tomich, J. M.; Venglarik, C. J.; Forman, H. J. *Am. J. Physiol. Lung Cell Mol. Physiol.* **2001**, *281*, L24−L30. (b) Jiang, C.; Lee, E. R.; Lane, M. B.; Xiao, X.-F.; Harris, D. J.; Cheng, S. H. *Am. J. Physiol. Lung Cell Mol. Physiol.* **2001**, *281*, L1164−L1172. (c) Rodgers H. C.; Knox, A. J. *Eur. Respir. J.* **2001**, *17*, 1314−1321.

(6) Scheerder, J.; van Duynhoven, J. P. M.; Engbersen, J. F. J.; Reinhoudt, D. N. *Angew. Chem., Int. Ed. Engl.* **1996**, *35*, 1090−1093.

(7) Koulov, A. V.; Mahoney, J. M.; Smith, B. D. *Org. Biomol. Chem.* **2003**, *1*, 27−29.

(8) Mahoney, J. M.; Stucker, K. A.; Jiang, H.; Carmichael, I.; Brinkmann, N. R.; Beatty, A. M.; Noll, B. C.; Smith, B. D. *J. Am. Chem. Soc.* **2005**, *127*, 2922−2928.

(9) Aldridge, S.; Fallis, I. A.; Howard, S. T. *Chem. Comm.* **2001**, 231−232.

(10) Andredakis, G. E.; Moschou1, E. A.; Matthaiou, K.; Froudakis, G. E.; Chaniotakis, N. A. *Anal. Chim. Acta* **2001**, *439*, 273−280.

(11) Pichierri, F. *J. Mol. Struct. (THEOCHEM)* **2002**, *581*, 117−127.

(12) Kim, S. K.; Singh, N. J.; Kim, S. J.; Kim, H. G.; Kim, J. K.; Lee, J. W.; Kim, K. S.; Yoon, J. *Org. Lett.* **2003**, *5*, 2083−2086.

(13) Yoon, J.; Kim, S. K.; Singh, N. J.; Lee, J. W.; Yang, Y. J.; Chellappan, K.; Kim, K. S. *J. Org. Chem.* **2004**, *69*, 581−583.

(14) Garau, C.; Quinonero, D.; Frontera, A., Costa, A.; Ballester, P.; Deya, P. M. *Chem. Phys. Lett.* **2003**, *370*, 7−13.

(15) Ruangpornvisuti, V. *J. Mol. Struct. (THEOCHEM)* **2004**, *686*, 47−55.

(16) Brynda, M.;,Tomasz A. Wesolowski, T. A.;Wojciechowski, K. *J. Phys. Chem. A* **2004**, *108*, 5091−5099.

(17) Vivas-Reyes, R.; De Proft, F.; Biesemans, M.; Willem, R.; Geerlings, P. *Eur. J. Inorg. Chem.* **2003**, 1315−1324.

(18) Mohamadi, F.; Richards, N. G. J.; Guida, W. C.; Liscamp, R.; Lipton, M.; Caufield, C.; Chang, G.; Hendrickson, T.; Still, W. C. *J. Comput. Chem.* **1990**, *11*, 440−449.

(19) Hout, R. F.; Levi, B. A.; Hehre, W. J. *J. Comput. Chem.* **1982**, *3*, 234−241.

(20) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; 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.: Pittsburgh, PA, 2003.

(21) Huber, K. P.; Herzberg, G. Molecular Spectra and Molecular Structure. In *Constants of Diatomic Molecules;* Van Nostrand: New York, 1979; Vol. IV.

(22) Garau, C.; Frontera, A.; Ballester, P.; Quinonero, D; Costa, A; Deya, P. M. *Eur. J. Org. Chem.* **2004**, *1*, 179−183.

(23) Sun, H; Kung, P. W. C. *J. Comput. Chem.* **2005**, *26*, 169−174.

(24) Bencivenni, L; Cesaro, S. N.; Pieretti, A. *Vib. Spectrosc.* **1998**, *18*, 91−102.

(25) (a) Buhl, M.; Ludwig, R.; Schurhammer, R.; Wipff, G. *J. Phys. Chem.* A **2004**, *108*, 11463−11468. (b) Anderson, J. D.; Paulsen, E. S., Dearden, D. V. *Int. J. Mass Spectr.* **2003**, *227*, 63−76. (c) Glendening, E. D.; Feller, D. *J. Am. Chem. Soc.* **1996**, *118*, 6052−6059.

(26) Wainwright, K. P. *Coord. Chem. Rev.* **1997**, *166*, 335−90.

(27) Menon, S. K.; Hirpara, S. V.; Harikrishnan, U. *Rev. Anal. Chem.* **2004**, *23*, 233−267.

(28) Wolfram Research, Inc. *Mathematica Version 5.2*; Wolfram Research, Inc.: Champaign, IL, 2005.

CT050270D

JCTC Journal of Chemical Theory and Computation

Design of Density Functionals by Combining the Method of Constraint Satisfaction with Parametrization for Thermochemistry, Thermochemical Kinetics, and Noncovalent Interactions

Yan Zhao, Nathan E. Schultz, and Donald G. Truhlar*

Department of Chemistry and Supercomputing Institute, University of Minnesota, 207 Pleasant Street S.E., Minneapolis, Minnesota 55455-0431

Received November 8, 2005

Abstract: We present a new hybrid meta exchange-correlation functional, called M05-2X, for thermochemistry, thermochemical kinetics, and noncovalent interactions. We also provide a full discussion of the new M05 functional, previously presented in a short communication. The M05 functional was parametrized including both metals and nonmetals, whereas M05-2X is a high-nonlocality functional with double the amount of nonlocal exchange (2X) that is parametrized only for nonmetals. In particular, M05 was parametrized against 35 data values, and M05-2X is parametrized against 34 data values. Both functionals, along with 28 other functionals, have been comparatively assessed against 234 data values: the MGAE109/3 main-group atomization energy database, the IP13/3 ionization potential database, the EA13/3 electron affinity database, the HTBH38/4 database of barrier height for hydrogen-transfer reactions, five noncovalent databases, two databases involving metal−metal and metal−ligand bond energies, a dipole moment database, a database of four alkyl bond dissociation energies of alkanes and ethers, and three total energies of one-electron systems. We also tested the new functionals and 12 others for eight hydrogen-bonding and stacking interaction energies in nucleobase pairs, and we tested M05 and M05-2X and 19 other functionals for the geometry, dipole moment, and binding energy of HCN−BF_3, which has recently been shown to be a very difficult case for density functional theory. We tested eight functionals for four more alkyl bond dissociation energies, and we tested 12 functionals for several additional bond energies with varying amounts of multireference character. On the basis of all the results for 256 data values in 18 databases in the present study, we recommend M05-2X, M05, PW6B95, PWB6K, and MPWB1K for general-purpose applications in thermochemistry, kinetics, and noncovalent interactions involving nonmetals and we recommend M05 for studies involving both metallic and nonmetallic elements. The M05 functional, essentially uniquely among the functionals with broad applicability to chemistry, also performs well not only for main-group thermochemistry and radical reaction barrier heights but also for transition-metal−transition-metal interactions. The M05-2X functional has the best performance for thermochemical kinetics, noncovalent interactions (especially weak interaction, hydrogen bonding, $\pi\cdots\pi$ stacking, and interactions energies of nucleobases), and alkyl bond dissociation energies and the best composite results for energetics, excluding metals.

1. Introduction

Kohn−Sham density functional theory (DFT) is now one of the most popular tools in the computational and theoretical chemistry community, and much progress has been made in the past decade in the development and validation of exchange and correlation functionals.[1−67] The line of research developing functionals by requiring them to satisfy constraints has led to the PW91,[4] PBE,[12] PKZB,[23] and TPSS[41] functionals on the second and third rungs of "Jacob's

* Corresponding author phone: (612) 624-7555; fax: (612) 624-9390; e-mail: truhlar@.umn.edu.

10.1021/ct0502763 CCC: $33.50
Published on Web 02/04/2006

ladder".[30] Although the PKZB functional proved disappointing,[23,26,44] the PBE and TPSS functionals have had some notable success in solid-state physics and some areas of chemistry.[42,44] However, as pointed out in a prescriptive paper by Perdew et al.,[61] PBE and TPSS are not suitable for kinetics (i.e., barrier heights) because both functionals seriously underestimate barrier heights; for example, they were found[55] to underestimate barrier height by an average of 8.5 kcal/mol for 76 barrier heights. The successful DFT methods for kinetics have been developed in a semiempirical way. This involves choosing a flexible functional form depending on one or more parameters and then fitting these parameters to a set of experimental or accurate data. MPW1K,[27] BB1K,[49] BMK,[50] MPWB1K,[51] and PWB6K[58] are examples of functionals for kinetics determined by the semiempirical approach. The semiempirical approach has also been used to obtain improved functionals for main-group thermochemistry, and a sequence of closely related papers leading successively to functionals called B97,[13] B98,[16] HCTH,[19] B97-1,[19] B97-2,[32] τ-HCTH,[34] τ-HCTHh,[34] BMK,[50] and B97-3[64] provides a good example of this approach. The successive functionals, however, may be improved for one kind of prediction but worsened for another, depending on changes in the functional form, optimization strategy, and training data. A common misconception is that the choice of training data is of overriding importance; actually, the choice of functional form is more critical in that, if the functional form is inadequate, one will not be able to fit a diverse set of data even if it is used for training. Nevertheless, the choice of data is sometimes critical as well. For example, BMK[50] is a functional using the same functional form as τ-HCTHh,[34] but it was reparametrized against a data set not only for thermochemistry but also for kinetics; the functional form and training set were well enough chosen that BMK performs equally well for kinetics and thermochemistry. However, BMK's performance for noncovalent interactions is inferior to, for example, PWB6K. PWB6K[58] has been shown to be a good functional for weak interactions, and it can describe stacking interactions in small organic clusters[59] and nucleobase pairs,[60] but its performance for thermochemistry is inferior to that of BMK. It has proved very challenging to develop a functional which can perform well for kinetics, main-group thermochemistry, and noncovalent interactions, including those in nonpolar weakly interacting systems and charge-transfer complexes.

It has been stated[42] that a "sophisticated nonempirical functional should provide a uniformly accurate description of diverse systems and properties, putting to rest the 'different functionals for different tasks' philosophy." Unfortunately, if one simultaneously considers metallic chemistry and barrier heights in open-shell systems, such a functional did not exist until, in a recent communication,[65] we reported a new functional, called M05, which was designed for very general purposes. The M05 functional performs well for all three of the properties mentioned at the end of the previous paragraph and also for transition-metal bond energies, ionization potentials (IPs), and electron affinities (EAs). One purpose of the present paper is to give a more complete account of this new functional. Another purpose is to present an alternative parametrization in which transition metals are not included in the training set. The new functional, to be called M05-2X, performs even better for kinetics, thermochemistry, and noncovalent interactions. Since a large number of important applications in chemistry and biochemistry do not involve transition metals, M05-2X may be very useful for such practical work. In contrast, the original M05 functional should be useful for problems involving bonds between two transition metals or metal−ligand bonds where one must treat general metals and organic or inorganic ligands accurately in the same system. In addition, the M05 functional has a fundamental importance in demonstrating the ability of a sufficiently flexible functional form containing kinetic energy density in both the exchange and correlation functionals and parametrized against a purposefully assembled and diverse data set to predict all the data reasonably well.

The M05 and M05-2X functionals belong to the fourth rung of Jacob's ladder (which is explained elsewhere[30,41]), and they, like the earlier B1B95,[10] τ-HCTHh,[34] TPSSh,[42] BB1K,[49] BMK,[50] MPW1B95,[51] MPWB1K,[51] PWB6K,[58] PW6B95,[58] and TPSS1KCIS functionals,[54] can be called hybrid metageneralized gradient approximations (hybrid meta-GGAs), because they incorporate electron spin density, density gradient, kinetic energy density, and Hartree−Fock (HF) exchange. Spin density, density gradient, and kinetic energy density are local properties of the density, although the latter two are sometimes called semilocal (in the early literature, they were sometimes incorrectly called nonlocal), whereas Hartree−Fock exchange is nonlocal. Including Hartree−Fock exchange is sometimes regarded as a temporary expedient that is necessary only because the local exchange-correlation functionals are insufficiently developed, but that is a misimpression. Perdew et al.[61] pointed out that, since the exact exchange energy of a fully spin-polarized one-electron system (like a hydrogen atom or H_2^+) is nonlocal, no local exchange-correlation functional can possibly be correct for this in general (of course, one could force any finite number of one-electron systems to be correct, but this is not the same as getting the effect exactly correct). Thus, the inclusion of Hartree−Fock exchange is a permanent feature of accurate exchange-correlation functionals, not a temporary expedient. The recent post-Hartree−Fock model[38,62,63] proposed by Becke employs 100% Hartree−Fock exchange. One line of argument would be that the ability to tolerate a high percentage of HF exchange and still give good results is the mark of a high-quality density functional.

For the present development effort, we combined the semiempirical approach with the incorporation of constraints in the new functionals. The constraints employed are as follows: (1) the new functionals are correct in the uniform electron gas (UEG) limit, and (2) the correlation functional should be free of self-interaction. The first condition is of fundamental importance, and any functional that violates the UEG limit cannot possibly be a universal functional. (Of course, any functional that gets the ionization potential of carbon wrong or the atomization energy (AE) of SiH_4 wrong also cannot be universal, but in the present article, the errors

in ionization potential and atomization energies are minimized with respect to parameter variations, whereas the UEG limit is actually constrained to be exact.) The second constraint is also important even though it does not remove the self-interaction error for the exchange part. Because we use both semiempirical parameter optimization and the method of constraint satisfaction, our approach may be considered to partake of key elements in both of the previously successful lines of functional development. Some workers make a distinction between fitting to analytic results such as fits to the artificial limit of a uniform electron gas or the analytic energy of a hydrogen atom and fitting to numerical results such as the energy of a helium atom, the ionization potential of carbon, or the hydrogen-bond strength of water dimer. Our own philosophy is to use both kinds of information for functional design. Another distinction sometimes made is between using parameters for fitting data and using parameters for shaping a functional. In designing a functional for broad applicability by not only incorporating constraints but also using training data, this distinction becomes arbitrary, and we will not be concerned with it.

Section 2 presents our training and test sets. Section 3 gives computational details. Section 4 discusses the theory and parametrization of the new functionals. Section 5 presents results and discusses them. Because the M05 functional was already discussed briefly in a preliminary communication,[65] we will discuss the M05-2X functional first.

2. Databases

2.1. M05-2X Training Set. The training set for the M05-2X models includes the six atomization energies in the AE6 representative database presented previously;[68] the binding energies of three dimers,[56] $(H_2O)_2$, $(CH_4)_2$, and $(C_2H_4)_2$; the binding energy[56] of the $C_2H_4\cdots F_2$ charge-transfer complex; the total atomic energies[69] of the H, C, O, S, and Si atoms; the ionization potentials[66] of C, O, OH, Cu, and Cr; the electron affinities[70] of C, O, and OH; the carbon−carbon bond dissociation energies[71] of the CH_3 bond with CH_3 and the isopropyl bond to CH_3; and the Kinetics9 database,[49,58] which is a database of three forward barrier heights, three reverse barrier heights, and three energies of reaction for the three reactions in the BH6[68] database. We have previously used Kinetics9 to optimize the BB1K,[49] MPWB1K,[51] and PWB6K[58] methods. Note that we used this small data set to parametrize the new methods, but we assess the new methods with several much larger data sets described below.

2.2. MGAE109/05 Test Set. The MGAE109/05 test set consists of 109 AEs for main-group compounds. All 109 data values are pure electronic energies; that is, zero-point energies and thermal vibrational−rotational energies have been removed by methods discussed previously.[54,70,72] The 109 molecules are part of Database/3,[72] and the atomization energies of NO, CCH, C_2F_4, and singlet and triplet CH_2 have been updated[54] recently. The updated data is a subset of Database/4.[73]

2.3. Ionization Potential and Electron Affinity Test Set. The zero-point-exclusive IP and EA test sets are called IP13/3 and EA13/3, respectively, and they are taken from a previous paper.[70] These data for six atoms and seven molecules are part of Database/3.

2.4. HTBH38/04 Database. The HTBH38/04 database contains 38 transition-state barrier heights for 19 hydrogen-transfer (HT) reactions, 18 of which involve radicals as reactants and products. They are taken from previous papers,[54,55] and they are also listed in the Supporting Information.

2.5. Noncovalent Interaction Databases. Recently, we developed several databases, in particular, HB6/04,[56] CT7/04,[56] DI6/04,[56] WI7/05,[58] and PPS5/05,[58] for various kinds of noncovalent interactions. HB6/04 is a hydrogen-bond database that consists of the equilibrium binding energies of six hydrogen-bonding dimers, namely, $(NH_3)_2$, $(HF)_2$, $(H_2O)_2$, $NH_3\cdots H_2O$, $(HCONH_2)_2$, and $(HCOOH)_2$. The CT7/04 database consists of the binding energies of seven charge-transfer complexes, in particular, $C_2H_4\cdots F_2$, $NH_3\cdots F_2$, $C_2H_2\cdots ClF$, $HCN\cdots ClF$, $NH_3\cdots Cl_2$, $H_2O\cdots ClF$, and $NH_3\cdots ClF$. The DI6/04 database contains the binding energies of six dipole interaction complexes: $(H_2S)_2$, $(HCl)_2$, $HCl\cdots H_2S$, $CH_3Cl\cdots HCl$, $CH_3SH\cdots HCN$, and $CH_3SH\cdots HCl$. The WI7/05 database consists of the binding energies of seven weak interaction complexes, namely, HeNe, HeAr, Ne_2, NeAr, $CH_4\cdots Ne$, $C_6H_6\cdots Ne$, and $(CH_4)_2$, all of which are bound by dispersion interactions. The PPS5/05 database consists of binding energies of five $\pi-\pi$ stacking complexes, namely, $(C_2H_2)_2$, $(C_2H_4)_2$, sandwich $(C_6H_6)_2$, T-shaped $(C_6H_6)_2$, and parallel-displaced $(C_6H_6)_2$.

2.6. Transition-Metal−Transition-Metal and Metal−Ligand Databases. We employ two databases involving metals. One[57] is for the atomization energies of transition-metal−transition-metal dimers, and it is called the TMAE4/05 database; it contains the bond energies of Cr_2, Cu_2, V_2, and Zr_2. The other,[66] called MLBE4/05, is for the metal−ligand bond energies in organometallic and inorganometallic complexes, and it contains the Cr−C, Ni−C, Fe−C, and V−S bond energies of $CrCH_3^+$, $NiCH_2^+$, $Fe(CO)_5$, and VS. These databases are representative subsets of the larger and more diverse TMAE9/05[57] and MLBE21/05[66] databases. In the present paper, we also use these databases to illustrate the performance of the M05 and M05-2X functionals for the energies of bonds involving metal atoms.

2.7. Alkyl Bond Dissociation Energy (ABDE) Database. This database contains four R−X bond dissociation energies D_e (R = Me and X = CH_3 and OCH_3). This is called the ABDE4/05 database. The reference D_0 values are taken from a recent paper by Izgorodina et al.,[71] and we used the B3LYP/6-31G(d) zero-point vibrational energies scaled with a scale factor of 0.9806[74] to obtain D_e.

2.8. Dipole Moment Database. This database consists of the fixed-geometry dipole moments for six molecules, namely, N6, H_2CO, CuH, BF, LiCl, and H_2O, where N6 is α-amino,ω-nitro-dodecahexaene, which has the formula $H_2N(CH{=}CH)_6NO_2$. This database is called the DM6/05 database. We use the MP2/6-31G geometry[75] for the N6 molecule, and the reference dipole moment is computed at the MP2/ 6-311+G(2df,2p) level of theory since previous work[76] showed good agreement between the MP2 and CCSD(T) levels of theory for a smaller basis set. For the

CuH molecule, we use the geometry from the modified coupled pair functional (MCPF) calculations of Langhoff and Bauschlicher.[77] The reference dipole moment for CuH is an average of the values (2.95 and 2.98 D, respectively) obtained by their MCPF calculation[77] and our own[65] CCSD(T)/ANO calculation, where ANO denotes the triple-ζ atomic natural orbital basis set of Widmark et al.[78,79] The geometries and accurate dipole moments for H_2CO, BF, LiCl, and H_2O are calculated at the CCSD(T)/aug-cc-pVTZ level of theory.

2.9. IPEA8 Database. The IPEA8 database contains the ionization potentials of C, O, OH, Cr, and Cu and the electron affinities of C, O, and OH.

2.10. AAE5 and AAE4 Databases. The AAE5 database consists of the total atomic energies[69] of H, C, O, S, and Si, and the AAE4 database is the same as AAE5 except that it excludes the atomic energy of H.

3. Computational Methods

3.1. Geometries, Basis Sets, and Spin−Orbit Energy. All calculations for the AE6, MGAE109/05, IP13/3, EA13/3, and HTBH38/04 databases are single-point calculations at QCISD/MG3 geometries, where QCISD is quadratic configuration interaction with single and double excitations[80] and MG3 is the modified[81,82] G3Large[83] basis set. The MG3 basis set,[81] also called G3LargeMP2,[82] is the same as 6-311++G(3d2f,2df,2p)[84,85] for H−Si but improved[83] for P−Ar.

The geometries for all of the molecules in the HB6/04, CT7/04, DI6/04, and WI7/05 noncovalent databases and the $(C_2H_4)_2$ and $(C_2H_2)_2$ dimers in the PPS5/05 database are optimized at the MC-QCISD/3 level, where MC-QCISD is the multicoefficient QCISD method.[72,86] The geometries for the benzene dimers in the PPS5/05 database are taken from Sinnokrot and Sherrill.[87]

The geometries for all of the molecules in the ABDE4/05 database are optimized at the B3LYP/6-31G(d) level, and they are taken from the Supporting Information of a previous paper.[71] The 6-311+G(3df,2p) basis set is used for the calculations of ABDEs for the purpose of comparison with the previous results.

The geometries for the molecules in the transition-metal−transition-metal (TMAE4/05) and metal−ligand (MLBE4/05) databases are optimized consistently with each level of theory. We used the double-ζ-quality DZQ basis set[57] for the calculations on the molecules in these two databases. The DZQ basis set uses the relativistic effective core potential method of Stevens et al.[88] for both the 3d and 4d transition metals, and it uses the 6-31+G(d,p) basis set for main-group atoms. In these cases (i.e., for the metal-compound calculations), the *d* functions are spherical harmonic 5D sets. Although one requires triple-ζ quality or better basis sets for quantitative results on transition metals, DZQ is good enough for a broad survey of many functionals to ascertain which ones gives relatively good results for bonds involving metal atoms.

The geometries for the stacked and hydrogen-bonded nucleobase pairs are optimized at the PWB6K/6-31+G(d,p) level. All DFT calculations for the base pairs use the 6-31+G(d,p) basis set.

To test the functionals for the one-electron systems, we employed the cc-pVQZ basis set for the hydrogen atom, H_2^+ (r_e = 1.4 bohr), and H_2^+ (r_e = 2.0 bohr). For the DM6/05 dipole moment database, we used the TZQ basis set, which is described in our previous paper.[57,66] For the MGAE109, HTBH38/04, IP13/3, EA13/3, and all five noncovalent databases, we used the MG3S basis sets for single-point energy calculations. The MG3S basis[70] is the same as MG3 except it omits diffuse functions on hydrogens.

Note that all of the basis sets mentioned above use pure *d* or *f* functions except the 6-31+G(d,p) basis set employed in the calculations for nucleobase pairs, which uses Cartesian basis functions.

In all of the calculations presented in this paper, the spin−orbit stabilization energy was added to atoms and open-shell molecules for which it is nonzero, as described previously.[57,66,81]

3.2. Counterpoise Correction. For noncovalent complexes, we perform calculations with and without the counterpoise corrections[89,90] for the basis set superposition error (BSSE).

3.3. Software. All of the calculations were performed with a locally modified version of the Gaussian03 program[91] except that the benchmark CCSD(T) calculations of the dipole moment for CuH were calculated with MOLPRO.[92]

4. Theory and Parametrization

4.1. Meta-GGA Exchange Functional. The functional form adopted for the meta-GGA exchange functional is

$$E_X^{(0)} = \sum_\sigma \int \mathrm{d}r\, F_{X\sigma}^{PBE}(\rho_\sigma,\nabla\rho_\sigma)\, f(w_\sigma) \tag{1}$$

where $F_{X\sigma}^{PBE}(\rho_\sigma,\nabla\rho_\sigma)$ is the exchange energy density of the PBE[11] exchange model (which has the same functional form as the earlier exchange functional of Becke,[1] but with different values for the two parameters) and $f(w_\sigma)$ is the kinetic-energy-density enhancement factor

$$f(w_\sigma) = \sum_{i=0}^{m} a_i w_\sigma^i \tag{2}$$

where the variable w_σ is a function of t_σ, and t_σ is a function of the kinetic energy density τ_σ of electrons with spin σ.

$$w_\sigma = (t_\sigma - 1)/(t_\sigma + 1) \tag{3}$$

where

$$t_\sigma = \tau_\sigma^{LSDA}/\tau_\sigma \tag{4}$$

$$\tau_\sigma = \frac{1}{2}\sum_i^{occup} |\nabla\Psi_{i\sigma}|^2 \tag{5}$$

$$\tau_\sigma^{LSDA} \equiv \frac{3}{10}(6\pi^2)^{2/3}\rho_\sigma^{5/3} \tag{6}$$

The motivation for the functional form in eqs 1−6 is explained in our previous paper,[65] and here we simply emphasize the key elements, namely, that it allows us to combine the correct UEG limit with reasonable behavior for

a large reduced density gradient and with Becke's strategy[18,25] for simulating delocalized exchange by local density functionals by using local functionals to detect delocalization and inhomogeneity.

4.2. Meta-GGA Correlation Functional. In the correlation functional, we treat the opposite-spin and parallel-spin correlations differently. We begin with Perdew and Wang's functional[5] for the correlation part of the local spin density approximation (LSDA). Then, following the analysis of Stoll et al.,[93] one can decompose the LSDA correlation energy into opposite-spin (denoted $\alpha\beta$) and parallel-spin (denoted $\sigma\sigma$, $\alpha\alpha$, and $\beta\beta$, depending on the content) correlation energy components for the UEG:

$$E_{C\alpha\beta}^{UEG}(\rho_\alpha,\rho_\beta) = E_C^{LSDA}(\rho_\alpha,\rho_\beta) - E_C^{LSDA}(\rho_\alpha,0) - E_C^{LSDA}(0,\rho_\beta) \quad (7)$$

$$E_{C\sigma\sigma}^{UEG}(\rho_\alpha) = E_C^{LSDA}(\rho_\alpha,0) \quad (8)$$

where $E_C^{LSDA}(\rho_\alpha,\rho_\beta)$ is the LSDA correlation energy. Recently, Gori-Giorgi et al.[94] showed that the spin resolution of the uniform electron gas correlation energy by eqs 7 and 8 is not accurate for spin-unpolarized ($\rho_\alpha = \rho_\beta$) systems. More recently, Gori-Giorgi and Perdew proposed a better formula.[95]

Note that eq 8 does not vanish in the one-electron case, and this nonvanishing is a manifestation of self-interaction error. To correct this self-interaction error, Becke[10] used a quantity, D_σ, which is defined as

$$D_\sigma = 2\tau_\sigma - \frac{1}{4}\frac{|\nabla\rho_\sigma|^2}{\rho_\sigma} \quad (9)$$

where τ_σ is the kinetic energy density of electrons with spin σ, defined in eq 10. The function D_σ can also be written as

$$D_\sigma = 2(\tau_\sigma - \tau_\sigma^W) \quad (10)$$

where τ_σ^W is the von Weizsäcker kinetic energy density[96] given by

$$\tau_\sigma^W = \frac{1}{8}\frac{|\nabla\rho_\sigma|^2}{\rho_\sigma} \quad (11)$$

In a one-electron case, $\tau_\sigma = \tau_\sigma^W$, so D_σ vanishes in any one-electron system. Note that the uniform electron gas limit ($\nabla\rho_\sigma \rightarrow 0$) of D_σ is

$$D_\sigma^{UEG} = \frac{3}{5}(6\pi^2)^{2/3}\rho_\sigma^{5/3} \quad (12)$$

Becke used D_σ/D_σ^{UEG} as a self-interaction correction factor to the parallel-spin case for the B95 correlation functional.[10] We have pointed out previously[56] that the function D_σ^{UEG} in the denominator causes some self-consistent field convergence problems that can be eliminated by using a different cutoff criterion. The D_σ^{UEG} in the denominator also causes some integration grid problems as pointed by Johnson and co-workers.[53] To avoid these numerical problems, we used a different self-interaction correction factor, $D_\sigma/2\tau_\sigma$ (also proposed by Becke[18]), which gives the right UEG limit but does not have the above-mentioned numerical instability.

The opposite-spins correlation energy of our new functional is expressed as

$$E_C^{\alpha\beta} = \int e_{\alpha\beta}^{UEG}\, g_{\alpha\beta}(x_\alpha,x_\beta)\, dr \quad (13)$$

where $g_{\alpha\beta}(x_\alpha,x_\beta)$ is defined as

$$g_{\alpha\beta}(x_\alpha,x_\beta) = \sum_{i=0}^{n} c_{C\alpha\beta,i}\left[\frac{\gamma_{C\alpha\beta}(x_\alpha^2 + x_\beta^2)}{1 + \gamma_{C\alpha\beta}(x_\alpha^2 + x_\beta^2)}\right]^i \quad (14a)$$

where

$$x_\sigma = \frac{|\nabla\rho_\sigma|}{\rho_\sigma^{4/3}} \qquad \sigma = \alpha, \beta \quad (14b)$$

For parallel spins,

$$E_C^{\sigma\sigma} = \int e_{\sigma\sigma}^{UEG}\, g_{\sigma\sigma}(x_\sigma)\frac{D_\sigma}{2\tau_\sigma}\, dr \quad (15)$$

where $D_\sigma/2\tau_\sigma$ is the self-interaction correction factor and

$$g_{\sigma\sigma}(x_\sigma) = \sum_{i=0}^{n} c_{C\sigma\sigma,i}\left(\frac{\gamma_{C\sigma\sigma}x_\alpha^2}{1 + \gamma_{C\sigma\sigma}x_\alpha^2}\right)^i \quad (16)$$

Note that $e_{\alpha\beta}^{UEG}$ and $e_{\sigma\sigma}^{UEG}$ in eq 13 and eq 15 are the UEG correlation energy density for the antiparallel-spin and parallel-spin cases, respectively, and they can be extracted from the total UEG correlation energy density in the same way as shown in eqs 7 and 8. The total correlation energy of the new correlation functional is given by

$$E_C = E_C^{\alpha\beta} + E_C^{\alpha\alpha} + E_C^{\beta\beta} \quad (17)$$

Note that our new correlation functional is similar to the correlation functional in the BMK[50] method; the difference is that BMK does not have the self-interaction correction factor $D_\sigma/2\tau_\sigma$ for the parallel-spin case.

We require $c_{C\alpha\beta,0} = c_{C\sigma\sigma,0} = 1$ in eqs 14a and 16. In agreement with the philosophy of the B95 functional,[10] this forces the correlation functionals to have the correct UEG limit, which is not enforced in a considerable body of work[16,19,34,50] using similar correlation functionals. One can easily confirm that our new correlation functional gives the right UEG limit (with $x_\sigma \rightarrow 0$, $D_\sigma \rightarrow 2\tau_\sigma \rightarrow D_\sigma^{UEG}$).

Following Becke,[18] we preoptimized the γ parameters to the correlation energies of He and Ne in a preliminary fit. The values of these two nonlinear parameters in the new functionals are

$$\gamma_{C\alpha\beta} = 0.0031 \text{ and } \gamma_{C\sigma\sigma} = 0.06 \quad (18)$$

4.3. Hybrid Meta Functional. The hybrid exchange-correlation energy can be written as follows:

$$E_{XC}^{hyb} = \frac{X}{100}E_X^{HF} + \left(1 - \frac{X}{100}\right)E_X^{DFT} + E_C^{DFT} \quad (19)$$

where E_X^{HF} is the nonlocal Hartree−Fock exchange energy,

Table 1. Optimized Parameters in the M05-2X and M05 Methods

	M05-2X			M05		
parameters	a_i	$c_{C\alpha\beta,i}$	$c_{C\sigma\sigma,i}$	a_i	$c_{C\alpha\beta,i}$	$c_{C\sigma\sigma,i}$
0	1.000 00	1.000 00	1.000 00	1.000 00	1.000 00	1.000 00
1	−0.568 33	1.092 97	−3.054 30	0.081 51	3.785 69	3.773 44
2	−1.300 57	−3.791 71	7.618 54	−0.439 56	−14.152 61	−26.044 63
3	5.500 70	2.828 10	1.476 65	−3.224 22	−7.465 89	30.699 13
4	9.064 02	−10.589 09	−11.923 65	2.018 19	17.944 91	−9.226 95
5	−32.210 75			8.794 31		
6	−23.732 98			−0.002 95		
7	70.229 96			9.820 29		
8	29.886 14			−4.823 51		
9	−60.257 78			−48.175 74		
10	−13.222 05			3.648 02		
11	15.236 94			34.022 48		
X		56			28	

X is the percentage of Hartree−Fock exchange in the hybrid functional, E_X^{DFT} is the local DFT exchange energy, and E_C^{DFT} is the local DFT correlation energy. Equation 19 can be rewritten as

$$E_{XC}^{hyb} = E_X^{HF} + \left(1 - \frac{X}{100}\right)(E_X^{DFT} - E_X^{HF}) + E_C^{DFT} \quad (20)$$

From eq 20, one can see that the total correlation energy for a DFT calculation is modeled as the sum of the dynamic correlation energy given by E_C^{DFT} and the nondynamical correlation energy[29] contained in $(1 - X/100)(E_X^{DFT} - E_X^{HF})$.

We optimize X along with the parameters in the new meta exchange and correlation functionals; the optimization procedure is given in the next section.

4.4. Optimization of the New Hybrid Meta-GGA. All of the parameter optimizations were carried out with a genetic algorithm.[97] The parameters a_i in eq 2 are determined by fitting them to the data in the training set with a constraint that $a_0 = 1$, which enforces the UEG limit. This limit corresponds to $t_\sigma = 1$, $w_\sigma = 0$, and $x_\sigma = 0$; and $f(w_\sigma)$ should tend to unity in this limit because the PBE exchange functional satisfies the UEG limit. Therefore, we constrained a_0 to unity to enforce this limit. Simultaneously, we optimized the $c_{C\alpha\beta,i}$ and $c_{C\sigma\sigma,i}$ parameters in eqs 14a and 16 to the data in the training set.

The M05 and M05-2X functionals were optimized using different training sets. In both new methods, we optimize the a_i parameters in the exchange functional, the $c_{C\alpha\beta,i}$ and $c_{C\sigma\sigma,i}$ parameters in the correlation functional, and the percentage, X, of Hartree−Fock exchange. We minimize the training function with respect to these parameters in a self-consistent way by solving the Fock−Kohn−Sham equation using the basis set and geometries described in section 3.1.

We optimized the parameters in M05-2X against the data in the training set to minimize the following training function

$$F = \text{RMSEPB(AE6)} + \text{RMSE(IPEA8)} + \text{RMSE(Kinetics9)} + 10 \times \text{RMSE(NB4)} + \text{RMSE(ABDE2)} + 0.2 \times \text{RMSE(AAE4)} + 2 \times \text{UE(AEH)} \quad (21)$$

where RMSEPB is the root-mean-squared error (RMSE) per bond. In particular, RMSEPB is obtained by dividing the RMSE for the AE6 database by the average number of bonds per molecule in this database. The second term is the RMSE for the IPEA8 database, which is defined in section 2. The third term is the RMSE for the Kinetics9 database. RMSE(NCCE4) is the RMSE for four noncovalent complexation energies, namely, the equilibrium binding energies of the $(H_2O)_2$, $(CH_4)_2$, and $(C_2H_4)_2$ dimers and that of the C_2H_4··· F_2 charge-transfer complex; RMSE(ABDE2) is the root-mean-square error in the bond dissociation energies of CH_3-CH_3 and isopropyl$-CH_3$; RMSE(AAE4) is the RMSE for the total atomic energies of C, O, S, and Si; and UE(AEH) is the unsigned error in the atomic energy of the hydrogen atom. Our preliminary fitting showed that, for the training function F, nonphysical parameters are produced when $m > 11$ or $n > 4$, so we used $m = 11$ in eq 2 and we used $n = 4$ in eqs 14a and 16. Thus, we optimized 20 parameters (11 in a_i, 4 in $c_{C\alpha\beta,i}$, and 4 in $c_{C\sigma\sigma,i}$ and X) for the M05-2X method.

All optimized parameters for M05-2X are listed in Table 1 along with the parameters for the M05 functional. In the optimization of the M05 functional,[65] the RMSE(ABDE2) is replaced by the RMSE for the bond dissociation energies of Cr_2 and V_2 and the Cr−C bond of $CrCH3^+$, and the weight we used for the error for the atomic energy of hydrogen is 0.2 instead of 2.

In the original work on the M05 functional, we found that we could obtain very similar results (the mean unsigned error for nonmetals was about 1% smaller and that for metals was about 13% larger) by employing the same strategy with the PBE exchange functional replaced by the mPW[15] one. Thus, the treatment of kinetic energy density and correlation energy, along with the consistency between the exchange and correlation functionals, is the key ingredients in the M05 functionals, not the precise form of $F_{x\sigma}$.

The optimized functions of eq 2 for the final M05 and M05-2X functionals are shown in Figure 1.

A useful way to visualize the meta-GGA nonlocality is to write the meta-GGA exchange-correlation energy as

$$E_{XC}[\rho_\alpha,\rho_\beta] = \int d^3r\, \rho\epsilon_X^{UEG}(\rho)\, F_{XC}(\rho_\alpha,\rho_\beta,\nabla\rho_\alpha,\nabla\rho_\beta,\tau_\alpha,\tau_\beta) \quad (22)$$

where $\rho = \rho_\alpha + \rho_\beta$ is the total density and $\epsilon_X^{UEG} = -(3/4\pi)(3\pi^2\rho)$ is the exchange energy per electron of a spin-unpolarized ($\rho_\alpha = \rho_\beta$) uniform electron gas; the enhancement factor F_{XC} shows the effects of correlation and inhomogeneity.[45] To visualize F_{XC} for the meta-GGA part of the M05 and M05-2X functionals, we define three quantities, namely,

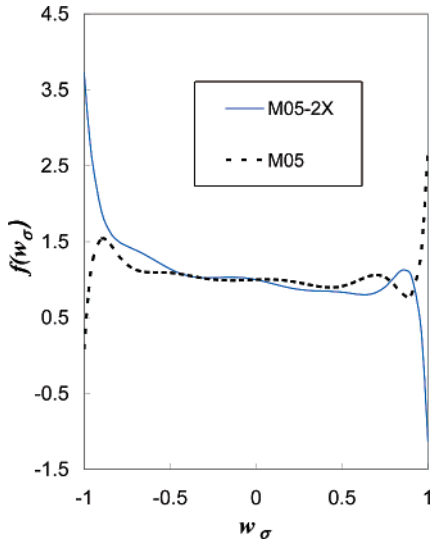


Figure 1. The τ enhancement factors for the M05 and M05-2X functionals.

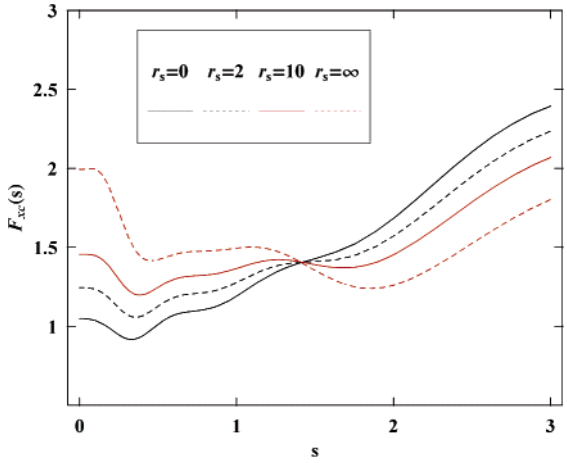


Figure 2. M05 enhancement factor F_{XC} of eq 22 as a function of the reduced gradient s of eq 23 with $\alpha_\tau = 0.2$ for various spin-unpolarized ($\rho_\alpha = \rho_\beta$) densities ranging from the high-density ($r_s = 0$) to the exchange-only limit ($r_s \rightarrow \infty$).

s, r_s, and α_τ

$$s = \frac{|\nabla\rho|}{(24\pi^2)^{1/3}\rho^{4/3}} \tag{23}$$

$$r_s = \left(\frac{3}{4\pi\rho}\right)^{1/3} \tag{24}$$

$$\alpha_\tau = \frac{\tau_\sigma - \tau_\sigma^W}{\tau_\sigma^{LSDA}} \tag{25}$$

By using eqs 6, 11, 23−25, we can transform the kinetic energy density into a function of s, r_s, and α_τ. Figures 2−5 show the enhancement factors for the meta-GGA part of the M05 and M05-2X functionals. Figures 2−5 show that both functionals violate the scaling inequality:[98]

$$F_{XC}(r'_s,s) > F_{XC}(r_s,s) \qquad r'_s > r_s \tag{26}$$

and they also violate the Lieb−Oxford bound[99]

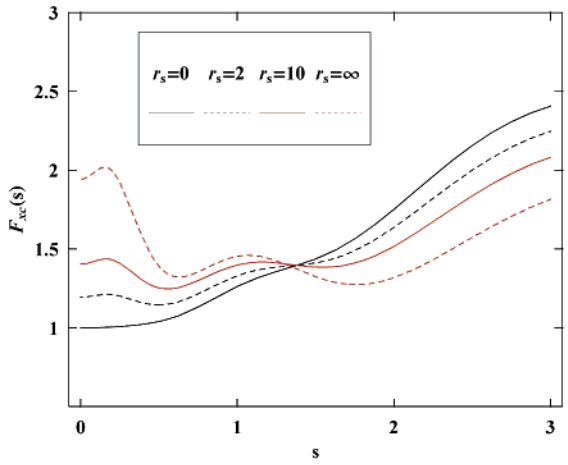


Figure 3. M05 enhancement factor F_{XC} of eq 22 as a function of the reduced gradient s of eq 23 with $\alpha_\tau = 1$ for various spin-unpolarized ($\rho_\alpha = \rho_\beta$) densities ranging from the high-density ($r_s = 0$) to the exchange-only limit ($r_s \rightarrow \infty$).

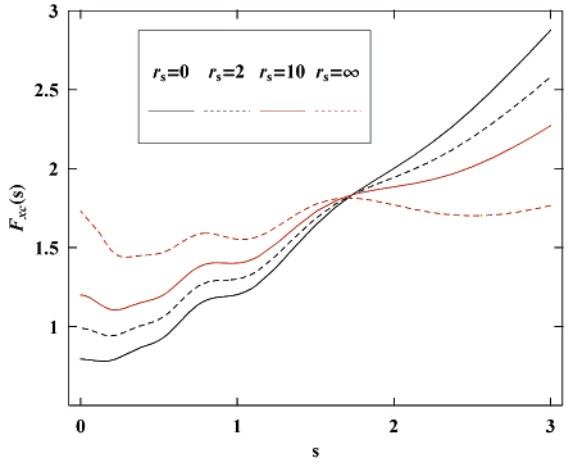


Figure 4. M05-2X enhancement factor F_{XC} of eq 22 as a function of the reduced gradient s of eq 23 with $\alpha_\tau = 0.2$ for various spin-unpolarized ($\rho_\alpha = \rho_\beta$) densities ranging from the high-density ($r_s = 0$) to the exchange-only limit ($r_s \rightarrow \infty$).

$$F_{XC}(r_s,s) \leq 2.273 \tag{27}$$

We note, though, that these figures only show the behavior of the meta-GGA part of the functional, and we do not recommend users to use the pure meta-GGA part of the M05 or M05-2X method, because the parameters of both functionals are optimized with the mixing of a certain amount of the Hartree−Fock exchange. We are working on the optimization of a pure local meta-GGA without the Hartree−Fock exchange using the same functional form.

5. Results and Discussion

5.1. Assessment of the New Hybrid Meta Functionals. We fitted our new functionals against a small and diverse data set (six data values in AE6, nine data values in Kinetics9, four data values for noncovalent complexation, eight data values for ionization energies and electron affinities, and three data values for transition-metal−transition-metal and metal−ligand interactions), but we assess the new functionals against a much larger data set that includes 109 main-group atomization energies, 13 ionization potentials (IP13), 13

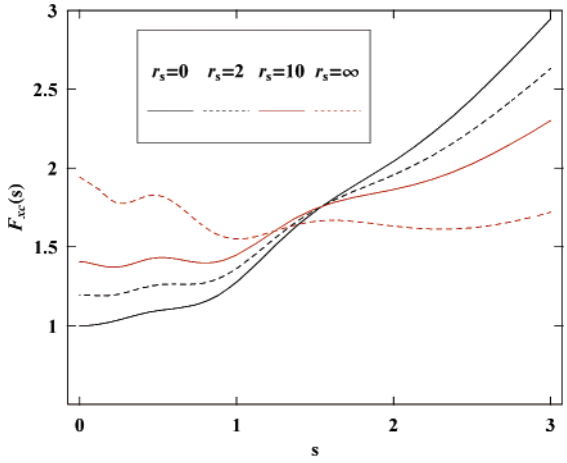


Figure 5. M05-2X enhancement factor F_{XC} of eq 22 as a function of the reduced gradient s of eq 23 with $\alpha_\tau = 1$ for various spin-unpolarized ($\rho_\alpha = \rho_\beta$) densities ranging from the high-density ($r_s = 0$) to the exchange-only limit ($r_s \rightarrow \infty$)

electron affinities (EA13), 38 barrier heights (HTBH38), 19 energies of reaction, 6 hydrogen-bonding energies (HB6), 7 charge-transfer complexation energies (CT7), 6 complexation energies of complexes dominated by dipole interaction (DI6), 7 weak interaction energies (WI7), 5 $\pi-\pi$ stacking interaction energies (PPS5), 4 transition-metal−transition-metal bond energies (TMAE4), 4 metal−ligand bond energies (MLBE4), 6 dipole moments, 4 alkyl bond dissociation energies ($R-CH_3$ and $R-OCH_3$), and 8 nucleobase pair interaction energies.

We compare the results obtained by the new methods to those for 28 other functionals. Table 2 lists all 30 density functionals considered in this work. In each case, we specify the year it was first published, the functional forms used for the dependence on $\nabla\rho$, whether the functional includes τ in the exchange and correlation functional, and whether the correlation functional is self-correlation-free (SCorF). Table 2 also contains two columns (one for the exchange functional and one for the correlation functional) that tell whether the functional reduces to the correct uniform electron gas limit when $\nabla\rho_\sigma \rightarrow 0$ and $\tau_\sigma \rightarrow \tau_\sigma^{\mathrm{LSDA}}$.

In most of the comparisons, we will gauge the quality of the results by mean unsigned errors (MUEs), which are the averages of the absolute deviations of the calculated values from database reference values, and by mean signed errors (MSEs), which are used to detect systematic deviations. However, for atomization energies, we use MUE per bond (MUEPB) and MSE per bond (MSEPB) because this allows[46,49] a more transferable comparison between databases with different average sizes of molecules. Because the dipole moments considered in the dipole moment database vary widely in magnitude, we also consider mean signed percentage error (MS%E) and mean unsigned percentage error (MU%E). We also use MU%E in one later table because the quantities considered in that table do not all have units of energy. To make the trends more clear, in every table, we will list the methods in increasing order of the values in the key (overall) error column, which is always the last column of a given table. The five smallest average errors for each of the individual databases and the five smallest average errors overall (for each table) are in bold.

5.2. Thermochemistry: AE, IP, and EA Results. Table 3 summarizes the errors in AEs, IPs, and EAs for all of the tested methods. Table 3 shows that the PW6B95, M05-2X, and BMK methods give the best results for AE calculations, and they give a MUEPB of less than 0.5 kcal/mol.

MPWB1K, BB1K, and MPW1B95 have the best performance for IP calculations, whereas BMK, PW6B95, and τ-HCTHh give the best performance for EA calculations.

To compare their performance for thermochemistry, we defined the TMUE (total MUE) as the mean signed error over all 135 data values in this table:

$$\mathrm{TMUE} = [\mathrm{MUEPB(AE)} \times 109 + \mathrm{MUE(IP)} \times 13 + \mathrm{MUE(EA)} \times 13]/135 \quad (28)$$

If we use TMUE as a criterion of practical usefulness for thermochemistry, Table 3 shows that M05-2X is the best functional, followed by PW6B95 and BMK.

5.3. Thermochemical Kinetics. Table 4 gives the mean errors for the HTBH38/04 database. A total of 18 of the 19 reactions in this database involve radicals as reactants or products, and 16 of those involve an odd number of electrons. Systems with an odd number of electrons and stretched bonds are well-known to provide a critical test case for density functional theory. Furthermore, we have shown elsewhere[55] that functionals that perform well for hydrogen-transfer barrier heights also perform well for barrier heights of more general classes of reaction, so we believe that good performance on this database is critical if a functional is to be judged as broadly applicable. Table 4 shows that BB1K gives the best results for barrier heights, with PWB6K, MPWB1K, MPW1K, BMK, and M05-2X being less accurate on average by 0.12−0.18 kcal/mol. M05 and the very new B97-3 are less accurate than these six functionals by 0.59−0.77 and 0.93−1.11 kcal/mol, respectively; whereas, the other 22 functionals in the table are less accurate than these six by 1.46−15.56 kcal/mol. The mean unsigned error in the energies of reaction for the 19 reactions is called MUE(ΔE19) and is given in the second-to-last column of Table 4. M05-2X gives the best performance for these energies of reaction, followed by B1B95, PW6B95, MPW1B95, BMK, and M05. Right behind these six are B98, τ-HCTHh, B3PW91, B97-2, and mPW1PW91, with the other 19 functionals being significantly less accurate. We also tabulated an average MUE (called AMUE) that is defined as

$$\mathrm{AMUE} = [\mathrm{MUE}(\Delta\mathrm{E19}) + \mathrm{MMUE(BH38)}]/2 \quad (29)$$

where MUE(ΔE19) is the mean unsigned error for the energy of reactions for the 19 reactions in the HTBH38 database. If we use AMUE as a criterion to judge the performance of a DFT method for thermochemical kinetics, Table 4 shows that M05-2X, BMK, BB1K, MPW1K, MPWB1K, and M05 are the best methods for kinetics. Table 4 is particularly encouraging in that M05-2X has a mean unsigned error for hydrogen-transfer barrier height on the order of 1.3 kcal/mol, a level of accuracy that is significantly exceeded only by the BB1K density functional, which is much less broadly applicable, and by large-basis CCSD(T) or some equally

Table 2. Density Functionals

			exchange				correlation			
method	year	ref(s)	$\nabla\rho$	X	τ?	UEG?	$\nabla\rho$	τ?	SCorF?	UEG?
BLYP	1988	2, 3	B88	0	no	yes	LYP	no	yes	no
SPWL	1992	5, 144	Slater	0	no	yes	PW91-L	no	no	yes
B3PW91	1993	2, 4, 7	B88	20	no	yes	PW91	no	no	yes
B3LYP	1994	2, 3, 8	B88	20	no	yes	LYP	no	yes	no
BB95	1996	2, 10	B88	0	no	yes	B95	yes	yes	yes
B1B95	1996	2, 10	B88	28	no	yes	B95	yes	yes	yes
G96LYP	1996	3, 9	G96	0	no	yes	LYP	no	yes	yes
PBE	1996	11	PBE	0	no	yes	PBE	no	no	yes
B1LYP	1997	2, 3, 14	B88	25	no	yes	LYP	no	yes	no
mPWPW91	1998	4, 15	mPW	0	no	yes	PW91	no	no	yes
mPW1PW91[a]	1998	4, 15	mPW	25	no	yes	PW91	no	no	yes
B98	1998	16	B98	21.98	no	no	B98	no	no	no
B97-1	1998	19	B97-1	21	no	no	B97-1	no	no	no
PBE1PBE[b]	1999	22	PBE	25	no	yes	PBE	no	no	yes
MPW1K	2000	27	mPW	42.8	no	yes	PW91	no	no	yes
B97-2	2001	19	B97-2	21	no	no	B97-2	no	no	no
τ-HCTHh	2002	34	τ-HCTHh	15	yes	no	τ-HCTHh	no	no	no
TPSS	2003	41	TPSS	0	yes	yes	TPSS	yes	yes	yes
TPSSh	2003	42	TPSS	10	yes	yes	TPSS	yes	yes	yes
X3LYP	2004	3, 47	X	21.8	no	yes	LYP	no	yes	no
BB1K	2004	2, 10, 49	B88	42	no	yes	B95	yes	yes	yes
BMK	2004	50	BMK	42	yes	no	BMK	no	no	no
MPW1B95	2004	10, 15, 51	mPW	31	no	yes	B95	yes	yes	yes
MPWB1K	2004	10, 15, 51	mPW	44	no	yes	B95	yes	yes	yes
TPSS1KCIS	2005	21, 41, 60	TPSS	13	yes	yes	KCIS	yes	yes	yes
PW6B95	2005	58	PW6B95	28	no	yes	PW6B95	yes	yes	yes
PWB6K	2005	58	PWB6K	46	no	yes	PWB6K	yes	yes	yes
B97-3	2005	64	B97-3	26.93	no	no	B97-3	no	no	no
M05	2005	65	M05	28	yes	yes	M05	yes	yes	yes
M05-2X	2005	present	M05-2X	56	yes	yes	M05-2X	yes	yes	yes

[a] Also called mPW0 and MPW25. [b] Also called PBE0.

Table 3. Mean Errors[a] [kcal/mol for Ionization Potentials (IP) and Electron Affinities (EA) and kcal/mol per Bond for Atomization Energies (AE)]

	MGAE109/05		IP13/3		EA13/3		
method	MSEPB	MUEPB	MSE	MUE	MSE	MUE	TMUE
PW6B95	−0.02	**0.40**	2.24	3.24	0.72	**1.78**	**0.81**
M05-2X	−0.02	**0.48**	1.69	3.54	0.53	2.03	**0.93**
BMK	−0.04	**0.47**	2.74	4.21	0.28	**1.56**	**0.94**
B1B95	−0.23	**0.55**	−0.13	**2.18**	3.02	3.16	**0.96**
MPW1B95	0.31	0.62	0.36	**2.14**	2.72	2.91	**0.98**
M05	−0.01	**0.53**	−0.41	2.87	2.81	2.96	0.99
B98	−0.50	0.64	1.99	3.21	0.30	**1.84**	1.00
B97-3	−0.37	0.59	1.56	3.51	0.82	2.07	1.02
B97-2	−0.20	0.65	0.46	**2.21**	2.41	2.89	1.02
TPSS1KCIS	−0.05	0.67	0.91	2.63	1.84	2.81	1.07
B97-1	−0.39	0.75	0.99	2.84	1.09	2.02	1.07
B3PW91	−0.13	0.66	3.70	4.25	−0.12	2.09	1.14
τ-HCTHh	−0.21	0.75	3.62	4.03	−1.18	**1.83**	1.17
PBE1PBE	0.11	0.91	2.44	3.23	1.50	2.76	1.31
mPW1PW91	−0.73	0.88	3.17	3.72	1.09	2.62	1.32
TPSS	0.63	1.03	1.80	3.11	0.51	2.31	1.36
TPSSh	−0.12	0.98	1.96	3.17	1.40	2.81	1.37
MPWB1K	−0.84	0.98	0.51	**2.05**	3.99	4.11	1.38
B3LYP	−0.69	0.91	3.58	4.72	−1.51	2.29	1.41
BB1K	−1.32	1.34	0.13	**2.09**	4.28	4.36	1.70
PWB6K	−1.41	1.43	1.57	2.28	3.23	3.59	1.72
X3LYP	−1.26	1.42	2.58	4.73	−0.41	3.04	1.89
BLYP	−0.47	1.49	−0.41	4.87	−0.11	2.63	1.93
mPWPW91	1.72	2.01	2.93	4.15	−1.56	2.26	2.24
G96LYP	−1.39	1.96	−1.12	4.64	1.33	2.93	2.31
BB95	2.18	2.34	−0.55	3.34	0.21	**1.99**	2.40
MPW1K	−2.33	2.34	3.41	3.53	2.79	3.71	2.59
B1LYP	−2.66	2.69	−0.13	3.80	2.56	3.64	2.89
PBE	2.80	3.03	2.11	3.58	−1.20	2.22	3.01
SPWL	16.89	16.89	4.34	5.18	−5.77	5.80	14.70
average[b]		1.68		3.43		2.77	1.95

[a] MUEPB denotes mean unsigned error (MUE) per bond. MSE denotes mean signed error. TMUE denotes total MUE, and it is defined as TMUE = [MUEPB × 109 + MUE(IP) × 13 + MUE(AE) × 13]/135. [b] In all tables, where the last row is "average", it is the average of that column for all functionals in the table.

Table 4. Mean Errors for Thermochemical Kinetics[a,b]

		HTBH38/04		ΔE19	
methods	X	MSE	MUE	MUE	AMUE[c]
M05-2X	56	−0.39	1.34	**0.64**	**0.99**
BMK	42	−0.82	**1.32**	**0.92**	**1.12**
BB1K	42	−0.57	**1.16**	1.38	**1.27**
MPW1K	42.8	−0.60	**1.32**	1.31	**1.32**
MPWB1K	44	−0.85	**1.29**	1.41	**1.35**
PWB6K	46	−0.50	**1.28**	1.57	1.42
M05	28	−1.20	1.93	0.95	1.44
B97-3	26.93	−2.11	2.27	1.15	1.71
B1B95	28	−2.80	2.80	**0.78**	1.79
MPW1B95	31	−3.02	3.02	**0.86**	1.94
PW6B95	28	−3.14	3.14	**0.85**	1.99
B97-2	21	−3.09	3.24	1.08	2.16
mPW1PW91	25	−3.54	3.55	1.13	2.34
B3PW91	20	−4.02	4.03	1.05	2.54
B98	21.98	−4.16	4.16	0.97	2.57
B1LYP	25	−2.84	3.18	2.29	2.73
PBE1PBE	25	−4.22	4.22	1.29	2.76
B97-1	21	−4.40	4.40	1.48	2.94
B3LYP	20	−4.13	4.23	1.95	3.09
τ-HCTHh	15	−5.29	5.29	0.97	3.13
TPSS1KCIS	13	−4.69	4.69	1.64	3.16
X3LYP	21.8	−3.98	4.09	3.03	3.56
G96LYP	0	−6.25	6.26	2.26	4.26
TPSSh	10	−5.97	5.97	2.65	4.31
BB95	0	−8.14	8.14	1.63	4.89
BLYP	0	−7.52	7.52	2.29	4.90
TPSS	0	−7.71	7.71	2.53	5.12
mPWPW91	0	−8.43	8.43	1.97	5.20
PBE	0	−9.32	9.32	2.71	6.01
SPWL	0	−17.72	17.72	6.39	12.05
Average			4.57	1.70	3.14

[a] The MG3S basis used for all calculations in this table. [b] MUE denotes mean unsigned error (kcal/mol). MSE denotes mean signed error (kcal/mol). [c] AMUE in this table is calculated by averaging the two MUE columns, and it is a measure of the quality of a method for kinetics.

expensive wave function theory. Furthermore, both M05-2X and M05 are in the top seven for each of these three mean unsigned error columns in Table 4. The only other functional that appears in the top-seven list for all three of these columns is BMK.

Table 5. Mean Errors for Noncovalent Databases (kcal/mol)[a,b,c]

	HB6/04			CT7/04			DI6/04			WI7/05			PPS5/05			
	MUE		MMUE	MUE		MMUE	MUE		MMUE	MUE		MMUE	MUE		MMUE	MMMUE
method	no-cp	cp		no-cp	cp		no-cp	cp		no-cp	cp		no-cp	cp		
M05-2X	0.40	0.20	**0.30**	0.46	0.30	**0.38**	0.27	0.32	**0.29**	0.09	0.03	**0.06**	0.49	0.71	**0.60**	**0.33**
PWB6K	0.44	0.34	**0.39**	0.25	0.16	**0.21**	0.24	0.32	**0.28**	0.15	0.07	**0.11**	0.79	1.00	**0.90**	**0.38**
M05	0.58	0.53	0.55	0.68	0.30	**0.49**	0.23	0.24	**0.23**	0.14	0.06	**0.10**	1.12	1.34	**1.23**	**0.52**
MPWB1K	0.41	0.70	0.56	0.24	0.45	**0.34**	0.50	0.65	0.57	0.08	0.16	0.12	1.32	1.57	1.45	**0.61**
PW6B95	0.53	0.78	0.65	0.69	0.47	0.58	0.40	0.49	0.45	0.11	0.09	**0.10**	1.21	1.44	**1.32**	**0.62**
MPW1B95	0.50	0.86	0.68	0.47	0.31	**0.39**	0.50	0.63	0.56	0.10	0.16	0.13	1.46	1.70	1.58	0.67
B97-1	0.45	0.45	**0.45**	1.17	0.89	1.03	0.28	0.30	**0.29**	0.10	0.11	**0.10**	1.57	1.78	1.68	0.71
PBE1PBE	0.40	0.28	**0.34**	1.04	0.75	0.90	0.35	0.38	**0.37**	0.12	0.18	0.15	1.84	2.09	1.96	0.74
B98	0.45	0.66	0.55	0.91	0.66	0.79	0.34	0.40	0.37	0.12	0.16	0.14	1.91	2.13	2.02	0.78
MPW1K	0.33	0.61	0.47	0.44	0.66	0.55	0.52	0.67	0.60	0.20	0.29	0.25	2.25	2.53	2.39	0.85
X3LYP	0.45	0.48	0.47	0.96	0.68	0.82	0.45	0.59	0.52	0.16	0.22	0.19	2.49	2.71	2.60	0.92
mPW1PW91	0.39	0.79	0.59	0.65	0.51	0.58	0.53	0.63	0.58	0.58	0.30	0.44	2.43	2.71	2.57	0.95
TPSS1KCIS	0.49	0.86	0.67	1.22	0.95	1.08	0.46	0.55	0.50	0.17	0.21	0.19	2.39	2.62	2.50	0.99
TPSSh	0.41	0.80	0.60	1.44	1.16	1.30	0.49	0.58	0.54	0.18	0.26	0.22	2.46	2.72	2.59	1.05
BMK	0.68	0.96	0.82	0.41	0.62	0.52	0.78	0.97	0.88	0.76	0.85	0.81	2.36	2.57	2.47	1.10
B3LYP	0.60	0.93	0.76	0.71	0.54	0.63	0.78	0.94	0.86	0.31	0.39	0.35	2.95	3.17	3.06	1.13
BB1K	0.99	1.37	1.18	0.68	1.00	0.84	1.02	1.16	1.09	0.34	0.44	0.39	2.03	2.27	2.15	1.13
PBE	0.45	0.32	**0.39**	2.95	2.63	2.79	0.46	0.40	0.43	0.13	0.15	0.14	1.86	2.09	1.97	1.14
B1LYP	0.72	1.05	0.88	0.49	0.45	0.47	0.93	1.09	1.01	0.30	0.39	0.35	3.06	3.27	3.16	1.17
B97−3	1.16	1.50	1.33	0.48	0.63	0.56	0.82	0.98	0.90	0.49	0.58	0.53	2.49	2.70	2.59	1.18
TPSS	0.45	0.82	0.63	2.20	1.86	2.03	0.52	0.56	0.54	0.19	0.26	0.22	2.53	2.78	2.66	1.22
B97-2	1.22	1.64	1.43	0.56	0.67	0.61	0.87	1.02	0.94	0.25	0.35	0.30	2.73	2.96	2.84	1.23
B1B95	1.31	1.69	1.50	0.53	0.72	0.62	1.11	1.26	1.19	0.42	0.51	0.47	2.34	2.58	2.46	1.25
mPWPW91	0.57	0.96	0.77	2.25	1.89	2.07	0.56	0.59	0.57	0.24	0.32	0.28	2.69	2.96	2.83	1.30
B3PW91	1.03	1.43	1.23	0.64	0.69	0.66	0.97	1.14	1.06	0.53	0.62	0.58	3.23	3.49	3.36	1.38
τ-HCTHh	1.94	2.58	2.26	1.60	1.42	1.51	0.75	1.01	0.88	0.44	0.33	0.38	2.11	2.37	2.24	1.45
BLYP	1.18	1.56	1.37	1.67	1.42	1.54	1.00	1.18	1.09	0.45	0.53	0.49	3.58	3.79	3.69	1.63
BB95	1.83	2.21	2.02	1.48	1.27	1.38	1.18	1.35	1.27	0.57	0.66	0.62	2.96	3.18	3.07	1.67
SPWL	3.13	2.67	2.90	5.61	5.23	5.42	2.16	1.95	2.05	0.20	0.10	0.15	0.35	0.43	**0.39**	2.18
G96LYP	2.95	3.30	3.13	1.20	1.28	1.24	2.57	2.74	2.65	1.37	1.47	1.42	5.19	5.41	5.30	2.75
average	0.88	1.11	1.00	1.14	1.02	1.08	0.73	0.84	0.79	0.31	0.34	0.33	2.21	2.43	2.32	1.10

[a] MUE denotes mean unsigned error (MUE). MMUE = [MUE(cp) + MUE(no-cp)]/2, and MMMUE = [MMUE(HB) + MMUE(CT) + MMUE(DI) + MMUE(WI) + MMUE(PPS)]/5; HB, hydrogen bonding; CT, charge transfer; DI, dipole interaction; WI, weak interaction; and PPS, $\pi-\pi$ stacking. [b] We use "no-cp" to denote the calculation without the counterpoise correction for the BSSE and use "cp" to denote the calculation with the counterpoise correction for the BSSE. [c] The MG3S basis set is used for calculations in this table.

5.4. Noncovalent Interactions. The mean errors for noncovalent interaction are listed in Table 5. In Table 5, we use "no-cp" to denote calculations without the counterpoise correction for the BSSE and we use "cp" to denote calculations that do include the counterpoise correction for the BSSE. In Table 5, we also defined a mean MUE:

$$\text{MMUE} = [\text{MUE(no-cp)} + \text{MUE(cp)}]/2 \qquad (30)$$

This is a reasonable error criterion because the cp correction is sometimes an overestimate of BSSE and because, in practical work, some calculations are carried out with cp corrections and some without.

Table 5 shows that PBE1PBE, M05-2X, PWB6K, and PBE give the best performance for calculating the binding energies of the hydrogen-bonding dimers in the HB6/04 database. Table 5 also shows that M05-2X, PWB6K, and M05 give a very good performance for calculating the binding energies for the complexes in the CT7/04 and DI6/04 databases. M05-2X, M05, and PW6B95 give the best performance for calculating the binding energies of the weak interaction complexes in the WI7/05 database.

We note that $\pi-\pi$ stacking interactions play a dominant role in stabilizing various biopolymers, for example, the double helix structure of DNA, and such interactions are also important for supramolecular design. Table 5 shows that the quality of M05-2X for describing $\pi-\pi$ stacking interactions is better than PWB6K. This is encouraging because we have already shown[59,60] that PWB6K performs unusually well for the stacking interactions in the small organic clusters and nucleobase pairs.

The overall performance for noncovalent interactions can be judged by the mean MMUE, which is defined as

$$\text{MMMUE} = [\text{MMUE(HB)} + \text{MMUE(CT)} + \text{MMUE(DI)} + \text{MMUE(WI)} + \text{MMUE(PPS)}]/5 \qquad (31)$$

Notice that the five component in eq 31 place different requirements on a density functional. For example, high accuracy for charge-transfer complexes is not well correlated with high accuracy for weak interactions. If we use MMMUE as a criterion to evaluate the overall performance of DFT methods for noncovalent interactions, we can see from Table 5 that M05-2X, PWB6K, M05, MPWB1K, and PW6B95 are the best functionals.

5.5. Composite Results for Main-Group Energetic Databases. Table 6 is a summary of the performance of the tested methods for thermochemistry, kinetics, and noncovalent interactions. The second-to-last column of Table 6 is an average of the three mean unsigned errors. The M05-2X functional has an average error 1.4 times smaller than that of the second best performing method (M05), followed by BMK, MPWB1K, and PW6B95.

We also computed a weighted average where each error is divided by the average error of all 30 functionals for that quantity; this is shown in the last column. With this scaled average, the M05-2X functional performs 1.4 times better than the second-best performing functional M05, followed by PW6B95, PWB6K, and MPWB1K.

Table 6. Composite Energetic Results (kcal/mol)

method	thermochemical TMUE	kinetics AMUE	noncovalent interaction MMMUE	average[a]	scaled average[b]
M05-2X	**0.93**	**0.99**	**0.33**	**0.75**	**0.36**
M05	0.99	1.44	**0.52**	**0.98**	**0.48**
PW6B95	**0.81**	1.99	**0.62**	**1.14**	**0.54**
PWB6K	1.72	1.42	**0.38**	1.18	**0.56**
MPWB1K	1.38	**1.35**	**0.61**	**1.11**	**0.56**
MPW1B95	**0.98**	1.94	0.67	1.20	0.58
BMK	**0.94**	**1.12**	1.10	**1.05**	0.61
B98	1.00	2.57	0.78	1.45	0.68
B97-1	1.07	2.94	0.71	1.57	0.71
B97-3	1.02	1.71	1.18	1.30	0.71
B1B95	**0.96**	1.79	1.25	1.33	0.73
PBE1PBE	1.31	2.76	0.74	1.60	0.74
mPW1PW91	1.32	2.34	0.95	1.54	0.76
BB1K	1.70	**1.27**	1.13	1.37	0.77
B97-2	1.02	2.16	1.23	1.47	0.77
TPSS1KCIS	1.07	3.16	0.99	1.74	0.82
MPW1K	2.59	**1.32**	0.85	1.58	0.84
B3PW91	1.14	2.54	1.38	1.69	0.88
B3LYP	1.41	3.09	1.13	1.87	0.91
τ-HCTHh	1.17	3.13	1.45	1.92	0.97
X3LYP	1.89	3.56	0.92	2.12	0.98
TPSSh	1.37	4.31	1.05	2.24	1.01
B1LYP	2.89	2.73	1.17	2.26	1.14
TPSS	1.36	5.12	1.22	2.57	1.14
mPWPW91	2.24	5.20	1.30	2.92	1.33
BLYP	1.93	4.90	1.63	2.82	1.35
BB95	2.40	4.89	1.67	2.99	1.43
PBE	3.01	6.01	1.14	3.39	1.50
G96LYP	2.31	4.26	2.75	3.11	1.68
SPWL	14.70	12.05	2.18	9.64	4.45
average	1.95	3.14	1.10	2.06	1.00

[a] (TMUE + AMUE + MMMUE)/3 in kcal/mol. [b] [(TMUE/1.95) + (AMUE/3.14) + (MMMUE/1.10)]/3; note that the scaled average is unitless.

5.6. Trends in Alkyl Bond Dissociation Energies. Recently, Izgorodina et al. reported a study of the performance of several DFT methods, for the prediction of absolute and relative R−X bond dissociation energies (BDEs) where R is an alkyl group (R = Me, Et, *i*-Pr, and *t*-Bu) and X is a substituent (X = H, CH_3, OCH_3, OH, and F), and they found that all of the tested DFT methods overestimate the stabilizing effect on BDEs in going from R = Me to R = *t*-Bu, leading in some cases to incorrect qualitative behavior. Note that their results are consistent with the trends for the reaction energies in Table 3 of an earlier paper by Dybala-Defratyka et al.[100] Some earlier studies by Curtiss et al.[101] had also shown that conventional DFT methods perform much worse for the enthalpies of formation of the larger molecules, and they concluded that this is due to a cumulative effect in the error for the larger molecules.

Table 7 summarizes the results for the trends in R−X BDEs (R = Me and *i*-Pr; X = CH_3 and OCH_3). Table 7 shows that M05-2X gives surprisingly good results for these BDEs; it gives a better performance than the expensive G3-RAD[102] method for the ABDE4/05 database, and it gives a MUE of only 0.6 kcal/mol, whereas BMK (the second-best DFT method) gives a MUE of 1.7 kcal/mol.

Eight functionals were tested against more data of this type, and the results are in Figures 6 and 7 and in the Supporting Information; the additional tests include larger alkyl groups than those present in ABDE4/05, but they yield similar conclusions to those drawn from Table 7. Figures 6 and 7 present the trends for eight relative R−X BDEs (R = Me, Et, *i*-Pr, and *t*-Bu; X = CH_3 and OCH_3), and both figures show that M05-2X, like other DFT methods but to a much lesser extent, tends to overestimate the BDE-lowering effect accompanying the increasing size of the alkyl group. In contrast, the wave-function-based method G3-RAD slightly underestimates the stabilizing effect on R−X BDEs

Table 7. Alkyl Bond Dissociation Energies (D_e, kcal/mol)[a,b]

method	R−CH_3 R = Me	R−CH_3 R = *i*-Pr	R−OCH_3 R = Me	R−OCH_3 R = *i*-Pr	MSE	MUE
exptl	97.39	95.00	89.79	91.51		
M05-2X	97.37	94.01	90.65	90.93	−0.18	**0.61**
G3-RAD[c,d]	96.91	94.95	90.53	92.87	0.39	**0.66**
BMK[d]	97.99	93.42	88.81	87.99	−1.37	**1.67**
MPW1B95[d]	98.90	92.78	88.79	86.68	−1.64	**2.39**
MPWB1K[d]	98.54	93.01	88.10	86.56	−1.87	**2.44**
PWB6K	97.96	92.64	87.48	86.21	−2.35	2.64
B1B95[d]	97.58	91.05	87.83	86.01	−2.80	2.90
BB1K	97.58	91.75	87.02	85.17	−3.04	3.14
PW6B95	97.26	91.10	87.28	85.18	−3.22	3.22
B97-1	97.45	90.83	87.05	84.29	−3.52	3.55
BB95	98.35	90.15	87.79	83.69	−3.43	3.91
PBE	96.79	89.65	87.24	84.08	−3.98	3.98
B97-2	97.68	90.18	86.77	83.10	−3.99	4.14
τ-HCTHh	96.51	89.51	86.93	83.78	−4.24	4.24
B97-3	96.76	89.78	85.86	82.78	−4.63	4.63
B98	95.73	89.10	86.06	83.31	−4.87	4.87
X3LYP	95.73	89.10	86.06	83.31	−4.87	4.87
PBE1PBE	95.23	89.29	85.63	83.59	−4.98	4.98
M05	94.47	86.99	86.32	82.77	−5.79	5.79
mPWPW91	94.58	87.22	85.26	81.87	−6.19	6.19
mPW1PW91[d]	93.28	87.16	84.37	82.86	−6.51	6.51
B3PW91	93.18	86.52	83.79	81.07	−7.28	7.28
MPW1K[d]	92.80	87.42	82.77	81.25	−7.36	7.36
TPSS1KCIS	92.11	85.56	83.07	80.44	−8.13	8.13
B3LYP[d]	91.58	85.01	82.58	80.06	−8.62	8.62
TPSSh	90.47	84.12	82.08	79.62	−9.35	9.35
TPSS	90.48	83.74	82.36	79.54	−9.39	9.39
B1LYP	89.73	83.44	80.46	78.25	−10.45	10.45
BLYP[d]	90.31	82.64	81.09	77.50	−10.53	10.53
G96LYP	89.01	80.68	79.64	75.40	−12.24	12.24
SPWL	115.56	108.51	108.10	105.49	15.99	15.99
average[e]						5.70

[a] The B3LYP/6-31G(d) geometries are used in all calculations in this table. [b] All DFT calculations in this table use the 6-311+G(3df,2p) basis set. [c] G3-RAD is the "Gaussian-3 for radicals" method of ref 102. [d] Data for these methods are taken from Izgorodina et al.[71] [e] Average excludes G3-RAD, which is a wave function method (not a density functional method).

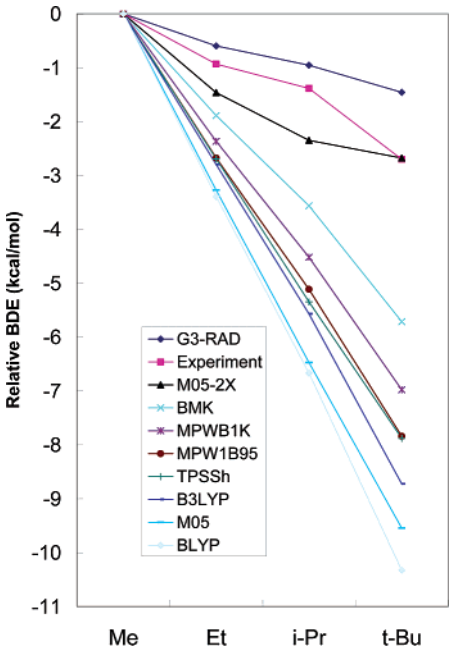


Figure 6. Effect of level of theory on the relative bond dissociation energies (in kcal/mol) for R−CH_3 species (R = methyl, ethyl, isopropyl, *tert*-butyl).

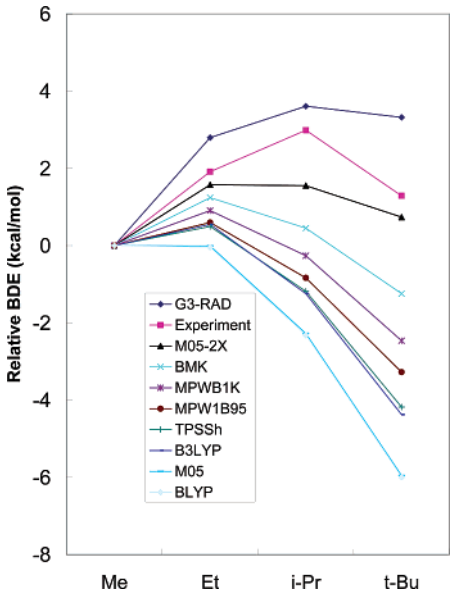


Figure 7. Effect of level of theory on the relative bond dissociation energies (in kcal/mol) for R−OCH_3 species (R = methyl, ethyl, isopropyl, *tert*-butyl).

on going from R = Me to *t*-Bu. Since Table 7 shows smaller mean unsigned errors for M05-2X than for G3-RAD, it is no longer appropriate to consider this kind of error as a failure of DFT, although it is a failure of some functionals.

The results in Table 7 and Figures 6 and 7 are encouraging because M05-2X shows small errors for the *absolute* and *relative* BDEs, and M05-2X offers promise as a reliable functional for larger systems.

5.7. Transition-Metal−Transition-Metal and Metal−Ligand Bond Energies. Metal−metal and metal−ligand bonding is very important in many application areas.[103−115] Table 8 summarizes the results for the TMAE4/05 and MLBE4/05 databases. For the TMAE4/05 database of bond energies of transition-metal dimers, BLYP, G96LYP, PBE, mPWPW91, and M05 give the best results. Note that M05-2X, PWB6K, MPWB1K, BB1K, and BMK are among the worst methods for transition-metal dimers because these DFT methods contain a large amount of HF exchange, and this makes the functionals less valid for systems with significant nondynamical correlation energy; hence, methods with correlation functionals that primarily account for dynamical (not static) correlation (this includes all 30 functionals tested in this article) and with more than 30% HF exchange are not recommended for studies of the interactions of transition-metal atoms with other transition-metal atoms where nondynamical correlation plays an important role; we will come back to this point in section 5.12.

For the MLBE4/05 database of metal−ligand compounds, TPSS1KCIS, TPSSh, M05, B97-2, and PBE1PBE give the best performance. In Table 8, MMUE is the average of the MUE for the TMAE4/05 and MLBE4/05 databases, and BLYP, M05, G96LYP, TPSS, and B97-2 give the smallest MMUEs. Notice that, of the 11 functionals with the smallest MMUEs, only M05 (X = 28) and B97-2 (X = 21) have X values larger than 15; six of these functionals have X = 0, and three (TPSSh, τ-HCTHh, and TPSS1KCIS) have X values in the range 10−15. The ability to obtain good results for bonds to metal atoms with an X value as large as 28 is one of the characteristics that allows M05 to have a broader range of applicability than any other functional.

Table 8. MUE (kcal/mol) for the TMAE4/05 and MLBE4/05 Databases with the DZQ Basis Set

	TMAE4/05		MLBE4/05		
method	MSE	MUE	MSE	MUE	MMUE[a]
BLYP	−0.86	**1.97**	9.23	9.23	**5.60**
M05	−5.98	**7.34**	−2.20	**4.97**	**6.15**
G96LYP	−5.71	**5.71**	6.99	8.10	**6.90**
TPSS	−6.18	8.38	7.00	7.00	**7.69**
B97-2	−10.07	10.95	−0.61	**5.52**	**8.24**
mPWPW91	−4.03	**7.28**	9.89	9.89	8.58
PBE	0.38	**5.87**	12.12	12.12	9.00
BB95	3.32	7.98	12.13	12.13	10.05
TPSSh	−15.97	15.97	1.42	**4.62**	10.30
τ-HCTHh	−5.91	13.07	3.56	7.68	10.38
TPSS1KCIS	−18.26	18.26	0.79	**4.39**	11.32
B97-1	−17.70	18.64	0.67	8.36	13.50
B3LYP	−21.47	21.47	−1.28	6.44	13.95
B98	−19.67	19.92	−0.73	8.00	13.96
X3LYP	−21.10	21.10	−1.75	6.86	13.98
B3PW91	−25.34	25.34	−2.54	**5.46**	15.40
PBE1PBE	−25.04	25.04	−3.34	6.31	15.68
PW6B95	−24.32	24.32	−4.00	7.36	15.84
B1B95	−25.13	25.13	−4.40	7.16	16.15
MPW1B95	−25.06	25.06	−4.64	7.61	16.33
B97-3	−22.80	22.80	−4.98	10.52	16.66
mPW1PW91	−26.46	26.46	−4.72	7.04	16.75
MPWB1K	−29.30	29.30	−11.35	11.52	20.41
BB1K	−29.56	29.56	−11.04	11.66	20.61
B1LYP	−27.14	27.14	−16.72	16.92	22.03
M05−2X	−21.92	29.42	−12.18	15.24	22.33
MPW1K	−31.83	31.83	−13.10	13.10	22.46
PWB6K	−33.90	33.90	−13.63	13.63	23.77
SPWL	23.03	23.03	30.30	30.30	26.66
BMK	−35.98	36.81	13.35	17.74	27.27
average		19.97		9.90	14.93

[a] MMUE = [MUE(TMAE4/05) + MUE(MLBE4/05)]/2.

5.8. Tests for Dipole Moments. Table 9 presents the performance for the DM6/05 database of the dipole moments. $NH_2(CH{=}CH)_6NO_2$ (denoted as N6) is a push−pull π-conjugated system, and the accurate evaluation of the electric dipole properties for this type of molecule is a difficult problem for density functional theory.[75,76] Among the tested DFT methods, M05-2X gives the best results for the dipole moment of N6, and in general, the DFT methods with higher percentages of Hartree−Fock exchange perform better than the DFT methods with lower (or zero) percentages of Hartree−Fock exchange. Overall, PWB6K gives the lowest MUE, followed by M05-2X, MPWB1K, BB1K, and BMK. If we consider the MU%E, MPW1B95 give the lowest MU%E, followed by PWB6K, M05, MPWB1K, and BB1K.

5.9. Tests for Noncovalent Interactions in Nucleobase Pairs. Table 10 summarizes the results for the stacking and hydrogen-bonding interactions in nucleobase pairs. All of the structures have been detailed in a previous paper,[60] and they are also given in the Supporting Information. For stacking interactions, SPWL and M05-2X give the best performance. However, the good performance of SPWL for stacking interactions is not matched by good accuracy for hydrogen bonding. Table 9 shows that SPWL gives the largest errors for the hydrogen-bonding interactions, while M05-2X gives the best performance for the interaction energies of the two Watson−Crick hydrogen-bonded base pairs.

The average MUE in Table 10 is defined as

$$\text{AMUE} = \text{MUE(stacking)} + \text{MUE(hydrogen bonding)} \quad (32)$$

Table 9. Dipole Moments Predicted by Density Functionals[a,b]

method	N6[c]	LiCl	H_2CO	CuH	H_2O	BF	MS%E[d]	MU%E[e]	MSE	MUE
accurate	11.56[f]	7.23[g]	2.39[g]	2.97[h]	1.85[g]	0.79[g]				
PWB6K	15.00	7.19	2.59	3.23	2.01	0.80	9.2	**9.4**	0.67	**0.69**
M05-2X	14.85	7.16	2.68	3.25	2.04	0.81	10.2	10.6	0.66	**0.69**
MPWB1K	15.07	7.19	2.58	3.20	2.01	0.81	9.3	**9.5**	0.68	**0.69**
BB1K	15.12	7.20	2.56	3.15	2.00	0.83	9.4	**9.5**	0.68	**0.69**
BMK	15.27	7.26	2.63	2.91	2.04	0.87	10.2	10.9	0.70	**0.72**
MPW1K	15.20	7.21	2.60	3.25	2.01	0.86	11.1	11.2	0.72	0.73
MPW1B95	15.55	7.13	2.49	2.92	1.98	0.85	8.2	**9.3**	0.69	0.74
B1B95	15.66	7.14	2.47	2.89	1.98	0.87	8.5	9.9	0.70	0.76
PW6B95	15.68	7.11	2.48	2.87	1.98	0.84	7.8	9.6	0.69	0.77
B97-3	15.76	7.17	2.50	2.90	1.98	0.92	10.2	11.3	0.74	0.79
PBE1PBE	15.85	7.11	2.47	2.88	1.97	0.92	9.8	11.4	0.74	0.81
mPW1PW91	15.87	7.13	2.49	2.89	1.98	0.91	9.7	11.2	0.74	0.81
M05	15.97	7.08	2.52	2.91	2.00	0.80	8.0	**9.5**	0.75	0.82
B97-2	16.04	7.14	2.46	2.91	1.96	0.91	9.9	11.1	0.77	0.82
B98	16.07	7.11	2.48	2.84	1.97	0.91	9.7	11.8	0.76	0.85
B97-1	16.06	7.10	2.46	2.80	1.96	0.92	9.5	12.0	0.75	0.85
B1LYP	15.94	7.08	2.51	2.77	1.98	0.88	8.7	11.7	0.73	0.85
B3PW91	16.11	7.12	2.46	2.79	1.97	0.93	9.8	12.3	0.76	0.86
X3LYP	16.10	7.06	2.49	2.70	1.98	0.89	8.6	12.4	0.74	0.89
B3LYP	16.18	7.07	2.48	2.68	1.98	0.90	8.6	12.7	0.75	0.90
TPSS1KCIS	16.38	7.08	2.42	2.76	1.95	0.94	9.5	12.6	0.79	0.91
τ-HCTHh	16.41	7.11	2.44	2.72	1.96	0.92	9.4	12.7	0.80	0.92
TPSSh	16.54	7.09	2.41	2.81	1.94	0.97	10.7	13.2	0.83	0.93
TPSS	17.06	7.03	2.34	2.60	1.92	1.01	10.2	16.0	0.86	1.07
mPWPW91	17.10	6.98	2.31	2.36	1.93	0.98	8.1	17.4	0.81	1.13
BB95	17.04	6.98	2.28	2.28	1.92	0.96	6.9	17.4	0.78	1.13
PBE	17.08	6.96	2.29	2.34	1.93	1.00	8.2	17.9	0.80	1.13
G96LYP	17.18	7.01	2.34	2.29	1.94	0.99	8.3	17.8	0.83	1.15
BLYP	17.17	6.94	2.33	2.25	1.93	0.96	7.3	17.6	0.80	1.16
SPWL	17.42	6.95	2.37	2.16	2.00	0.99	8.6	19.4	0.85	1.22
average								12.6		0.88

[a] All values are in Debyes. [b] All DFT calculations are single-point calculations using the TZQ basis set. [c] $NH_2(CH{=}CH)_6NO_2$ is denoted as N6. [d] Mean percentage signed error. [e] Mean percentage unsigned error. [f] MP2/6-311+G(2df,2p) result. All calculations use MP2/6-31G geometry for this molecule. [g] CCSD(T)/aug-cc-pVTZ result. All calculations use CCSD(T)/aug-cc-pVTZ geometry for these molecules. [h] The reference dipole moment for CuH is an average of the values by the MCPF calculation and a CCSD(T)/ANO calculation performed in the present study, where ANO is the triple-ζ atomic natural orbital basis set of Widmark et al. The geometry is taken from a previous study by Langhoff and Bauschlicher, and all calculations use this geometry ($r_{Cu-H} = 1.509$ Å).

Table 10. Results for Stacking and Hydrogen-Bonding Interactions in Nucleobase Pairs (kcal/mol)

	stacking								hydrogen bonding				
methods	A···T S	G···C S	C···C AP[a]	C···C D[a]	C···C S[a]	U···U S	MSE	MUE	A···T WC	G···C WC	MSE	MUE	AMUE[b]
best estimate[c]	11.60	16.90	9.90	9.43	−2.45	10.30			15.40	28.80			
M05-2X	10.28	16.25	10.52	10.02	−5.08	8.76	−0.82	1.22	14.56	28.58	−0.53	0.53	**0.87**
PWB6K	9.50	14.86	10.88	9.66	−5.93	7.94	−1.46	1.86	14.22	28.39	−0.79	0.79	**1.33**
MPWB1K	8.19	13.68	9.63	8.63	−7.02	6.51	−2.68	2.68	13.42	27.45	−1.67	1.67	**2.17**
PW6B95	7.68	13.10	9.48	8.46	−6.56	6.45	−2.84	2.84	13.26	26.68	−2.13	2.13	**2.49**
MPW1B95	7.47	12.83	8.98	8.08	−7.10	6.01	−3.24	3.24	13.18	26.80	−2.11	2.11	**2.67**
M05	5.77	11.95	7.86	7.66	−6.40	5.79	−3.84	3.84	13.68	27.07	−1.72	1.72	2.78
PBE1PBE	3.54	10.44	5.30	5.81	−8.40	3.97	−5.84	5.84	14.42	28.43	−0.67	0.67	3.25
B97-1	3.54	10.26	5.64	6.31	−8.01	4.05	−5.65	5.65	14.08	27.44	−1.34	1.34	3.50
BMK	5.42	11.54	6.37	6.17	−9.14	4.42	−5.15	5.15	12.49	26.30	−2.71	2.71	3.93
τ-HCTHh	2.39	9.38	4.54	5.14	−8.87	3.05	−6.68	6.68	13.79	27.42	−1.49	1.49	4.08
TPSSh	1.42	8.47	3.75	4.46	−9.64	2.29	−7.49	7.49	13.37	26.80	−2.02	2.02	4.75
SPWL	12.59	18.90	11.62	10.70	−2.33	10.64	1.07	1.07	22.30	39.44	8.77	8.77	4.92
B3LYP	−0.10	7.39	2.87	3.64	−10.70	1.47	−8.52	8.52	12.73	26.17	−2.65	2.65	5.59
B97-3	0.71	7.60	3.45	4.24	−10.21	1.90	−8.00	8.00	11.78	24.78	−3.82	3.82	5.91
average								4.58				2.32	3.45

[a] 6-31+G(d,p) is used for all calculations in this table. [b] AP denotes antiparallel, D denotes displaced, and S denotes sandwich. The structures for all base pairs in this table can be found in ref 60 and in the Supporting Information. [c] See ref 60 and references therein for the sources of these best estimates. [d] AMUE = 0.5MUE(Stacking) + 0.5MUE(Hydrogen Bonding).

M05-2X gives the lowest AMUE, followed by PWB6K and MPWB1K.

5.10. Tests for One-Electron Systems. Table 11 presents the results for three one-electron systems, namely, the hydrogen atom, H_2^+ with a bond distance of 1.4 bohr, and H_2^+ with a bond distance of 2.0 bohr. In Table 11, PWB6K gives the lowest MUE, followed by BB1K and BMK. All of the mean errors are disconcertingly large, but it is encouraging that the better functionals have errors 1.5−3 times smaller than those of the popular B3LYP. Again, the DFT methods with higher percentages of Hartree−Fock exchange generally (but not always) perform better than the DFT methods with lower (or zero) percentages of Hartree−Fock exchange.

It is interesting to note that five of the six best functionals for thermochemical kinetics (Table 4) are also among the six best density functionals in Table 11. B1B95 and MPW1B95 also rank in the top 10 of both tables. However, one cannot generalize this result because there are also cases where the performance in these two tables does not correlate.

Table 11. Predicted Energy for H and H_2^+[a]

methods	energy (hartree)			mean errors (kcal/mol)	
	H	H_2^+ (1.4 b)	H_2^+ (2.0 b)	MSE	MUE
HF	−0.499946	−0.569830	−0.602521	0 [b]	0[b]
PWB6K	−0.500452	−0.570071	−0.605866	−0.86	**0.86**
BB1K	−0.498539	−0.568421	−0.604787	0.11	**1.06**
BMK	−0.498903	−0.567576	−0.604547	0.27	**1.11**
X3LYP	−0.499785	−0.569126	−0.607086	−0.77	**1.14**
MPWB1K	−0.497995	−0.567491	−0.603772	0.64	**1.16**
M05-2X	−0.499743	−0.571127	−0.607252	−1.22	1.30
B1B95	−0.498260	−0.568153	−0.605486	0.08	1.32
B1LYP	−0.498204	−0.568099	−0.605644	0.07	1.38
PBE	−0.499854	−0.569849	−0.609222	−1.39	1.42
MPW1B95	−0.497603	−0.567017	−0.604200	0.73	1.43
M05	−0.497839	−0.570464	−0.607749	−0.79	1.67
G96LYP	−0.499052	−0.570323	−0.609532	−1.38	1.76
BB95	−0.497781	−0.567708	−0.607079	−0.06	1.85
BLYP	−0.497781	−0.567708	−0.607079	−0.06	1.85
PW6B95	−0.501499	−0.571477	−0.608338	−1.89	1.89
TPSS1KCIS	−0.500036	−0.572567	−0.609298	−2.01	2.01
PBE1PBE	−0.501227	−0.571595	−0.609083	−2.01	2.01
TPSSh	−0.500043	−0.572672	−0.609564	−2.09	2.09
TPSS	−0.500069	−0.573028	−0.610440	−2.35	2.35
B3LYP	−0.502346	−0.572079	−0.610047	−2.55	2.55
B98	−0.502865	−0.574646	−0.612113	−3.62	3.62
mPWPW91	−0.503098	−0.574019	−0.612966	−3.72	3.72
B97-1	−0.502785	−0.574955	−0.612360	−3.72	3.72
mPW1PW91	−0.503839	−0.574884	−0.612020	−3.86	3.86
MPW1K	−0.504420	−0.575563	−0.611473	−4.01	4.01
B3PW91	−0.504154	−0.575088	−0.612744	−4.12	4.12
B97-3	−0.503829	−0.575986	−0.613022	−4.30	4.30
B97-2	−0.504206	−0.578058	−0.615081	−5.24	5.24
τ-HCTHh	−0.507268	−0.580641	−0.618280	−7.09	7.09
SPWL	−0.478593	−0.540711	−0.583762	14.48	14.48
average					2.88

[a] The cc-pVQZ basis set is employed in all calculations in this table. [b] For a one-electron system, Hartree−Fock is the same as full configuration interaction for a given basis set, and the error in the density functional calculations are computed relative to these results.

For example, MPW1K and B97-3 rank much higher in Table 4 than in Table 11, and X3LYP and PBE rank much higher in Table 11 than in Table 4.

5.11. Tests for a Donor−Acceptor System: HCN−BF3. Recently, Philips and Cramer[116] reported a study of a boron−nitrogen complex, namely, HCN−BF3.[117,118] This is an example of a Lewis acid−base complex, also called a dative bond or a coordinate covalent bond. They employed 12 GGA and hybrid GGA functionals as well as some wave-function-based methods, and their conclusion was that "all DFT methods fail to predict a binding energy that compares favorably to the MCG3//MC-QCISD result of −5.7 kcal/mol." In particular, all DFT methods tested gave bond energies in the range 1.8−4.3 kcal/mol, except MPW1K and BLYP, which respectively yielded 4.7 and 7.4 kcal/mol. Table 12 shows the results for the 14 DFT methods tested in Table 10; of these, 13 were not tested in Philips and Cramer's paper. Table 12 also includes the B3PW91 method, which was judged[116] the overall best for structures and frequencies; MPW1K, which was the best (of those functionals tested[116]) for complexation energy; and PBE, TPSS, BB95, B1B95, and BB1K, which are added for their fundamental interest and to illustrate the dependence on the fraction of Hartree−Fock exchange. Table 12 shows that the binding energies calculated by the PWB6K, MPWB1K, and M05-2X methods agree well with the best estimate (MCG3//MC-QCISD/3 calculation), and M05-2X even predicts more attraction than MCG3//MC-QCISD/3. It is encouraging that M05 is the most successful functional for this system and that Table 12 shows that almost all of the functionals in Table 10 are better on average than B3PW91; M05 and PW6B95 are more accurate, on average, by a factor of 2.9 and 2.5, respectively, as well as being more accurate than MP2. M05-2X reduces the error in B3PW91, on average, by 33%.

Table 12. Bond Length, Dipole Moment, and Binding Energy of HCN−BF_3[a]

methods	R_{BN} (Å)	μ (D)	ΔE (kcal/mol)	M%UE[b]
best estimate	2.473[c]	4.14[d]	−5.7[e]	
M05	2.492	4.30	−5.0	5.6
PW6B95	2.427	4.39	−5.0	6.5
MP2	2.361	4.51	−6.3	8.0
B97-1	2.500	4.29	−4.6	8.1
PWB6K	2.292	4.77	−5.9	8.7
MPW1B95	2.348	4.60	−4.9	10.2
M05-2X	2.352	4.73	−6.5	11.3
PBE1PBE	2.348	4.64	−4.8	11.1
MPWB1K	2.253	4.89	−5.3	11.3
PBE	2.407	4.45	−4.3	11.4
τ-HCTHh	2.426	4.46	−4.2	11.9
B3LYP	2.535	4.23	−3.8	12.7
MPW1K	2.323	4.74	−4.7	12.7
BB1K	2.346	4.63	−4.5	12.9
BMK	2.351	4.72	−4.5	13.3
B1B95	2.432	4.39	−3.8	13.6
B97-3	2.615	4.09	−3.6	14.6
B3PW91	2.465	4.37	−3.2	16.6
TPSSh	2.230	4.98	−4.3	18.1
TPSS	2.239	4.93	−4.1	18.8
BB95	2.538	4.12	−2.7	18.6
SPWL	1.731	6.97	−12.1	70.4
average[f]	2.364	4.65	−4.9	14.8

[a] All DFT and MP2 results are for the aug-cc-pVTZ basis set. [b] Mean percentage unsigned error. [c] Experimental result.[117] [d] Experimental result.[118] [e] MCG3//MCQCISD/3 result.[116] [f] Average excludes MP2, which is a wave function method, not a density functional method.

5.12. Multireference Character. A simple and useful way to describe the optimum domains of applicability of the M05 and M05-2X functionals is that the former is recommended

for systems containing metals or transition metals (especially those in groups 2−10) and the latter is recommended for systems containing only nonmetallic or only main-group elements (although M05 is also very good for such systems, as shown in Tables 6, 8, and 10). This way of classifying systems, though, does not really capture the essence of the issue at a higher level of sophistication. We believe that the essential distinction is multireference character. Systems with significant multireference character are not well described by most density functionals that have more than 5−15% Hartree−Fock exchange. A system with large multireference character is one for which no single configuration-state function provides a good zero-order description;[119,120] such a system is said to contain significant amounts of static, near-degeneracy, or nondynamical correlation energy, often associated with multicenter systems, but also found in atoms.[121] Having made this distinction, one might summarize the situation as follows: M05 is recommended for applications where the systems studied involve both multireference and single-reference behavior, whereas if only single-reference behavior is to be encountered, one can obtain higher quantitative accuracy by switching to M05-2X. This is more satisfactory than the formulation at the start of this paragraph, but only partly more satisfactory because "multireference character" is not completely unambiguous.

One can characterize multireference character by analyzing a configuration interaction[122] or coupled cluster calculation,[123] but this is often impractical. In a recent paper,[66] we proposed a simpler criterion for the multireference character of a bond. We called this the B_1 diagnostic and defined it as

$$B_1 = [D_e(\mathrm{BLYP}) - D_e(\mathrm{B1LYP//BLYP})]/n \qquad (33)$$

where D_e is the energy required to break n bonds and B1LYP//BLYP denotes a B1LYP calculation of the same quantity using the BLYP equilibrium geometries for the molecule and the fragments. For B1LYP, the percentage of Hartree−Fock exchange is 25. The B_1 diagnostic measures multireference character because the Hartree−Fock exchange approximation fails badly for multireference systems, whereas GGAs can usually handle these systems almost as well as they handle single-reference systems. We previously concluded that bonds with $B_1 \lesssim 10$ kcal/mol are reasonably classified as single-reference cases, whereas those with $B_1 \gtrsim 10$ kcal/mol should be classified as multireference. This criterion is clearly not sophisticated enough to supplant system-specific discussions of metallic and multireference character,[77,120,121,124−133] and it does not fully supersede characterizing bonds in chemical terms,[66] but its ease of use is appealing.

In Tables 13 and 14, we present results for eight systems, four of which have B_1 values less than 10 kcal/mol (single reference) and four of which have B_1 values greater than 10 kcal/mol (multireference). Ozone (O_3) is a well-studied multireference system,[122] and its B_1 is about 22 kcal/mol. The cases in Table 13 were chosen so that two of the single-reference cases involve transition metals, one involves a main-group metal, and one has no metals. Similarly, two of the multireference cases involve metallic elements and two involve only nonmetals. Table 14 shows that the quality of the predictions depends more on multireference character than on the metallic character. For the prediction of the bond energies in systems with a low B_1 value, M05 and M05-2X perform equally well, but for the systems with a high B_1 value, M05-2X performs much worse than M05.

Table 13. Dissociation Energies (kcal/mol) and B_1 Values for Eight Bond-Breaking Processes[a]

	D_e			
process	experiment[b]	BLYP	B1LYP//BLYP	B_1
$CH_4 \rightarrow CH_3 + H$	112.7	109.9	109.5	0.4
$LiCl \rightarrow Li + Cl$	113.9	108.2	107.4	0.8
$AgCu \rightarrow Ag + Cu$	40.9	41.7	36.2	5.5
$Cu_2 \rightarrow 2Cu$	47.2	46.4	39.7	6.7
$VS \rightarrow V + S$	106.9	111.1	93.7	17.4
$CN \rightarrow C + N$	180.6	190.9	173.0	17.9
$O_3 \rightarrow 3O$	146.1	170.1	126.3	21.9[c]
$ZrV \rightarrow Zr + V$	61.9	72.7	32.9	39.8

[a] The TZQ basis set is used. The TZQ basis always uses spherical harmonic *d* and *f* functions (5D 7F sets). [b] The experimental values for CH_3-H and CN are calculated by using the experimental atomization energies from Database/3.[70] The experimental value for O_3 is taken from Database4/05.[73] The experimental values for Cu_2 and AgCu are taken from a previous paper,[57] and the experimental values for VS and ZrV are taken from a recent paper.[66] All dissociation energies in this table are zero-point-exclusive and spin−orbit-inclusive. [c] We put $n = 2$ in eq 33 in this case because two bonds are broken (not counting the long "bond").

Note that eq 33 does not apply to transition states, but the reader should be aware that transition states, even those for radical reactions, are not all multireference systems, although it is a common misconception that they are. For example, a multireference plus single and double excitation calculation lowers the barrier height of the $H + H_2$ reactions by only 0.3 kcal/mol as compared to a single-reference calculation with single and double excitations.[134] Similarly, single-reference plus dynamical-correlation-energy treatments give reasonable descriptions of the F−H−H and H−F−H transition states.[135] These conclusions based on wave function theory are consistent with our DFT findings that several methods fail quite badly for multireference systems with $B_1 > 10$ kcal/mol but are nevertheless quite accurate for transition states, even radical transition states. Examples would be MPW1K, BB1K, and PWB6K. In this light, the good performance of the M05 functional both for $B_1 > 10$ kcal/mol systems and for barrier heights is even more dramatic.

A final comment on transition-metal systems is warranted. In particular, it should be noted that complexes in which a transition metal is saturated with ligands or is one ligand short of saturation may have far less multireference character than highly unsaturated systems such as metal-containing diatomic molecules. Thus, the cases in our metallic training sets are more difficult than the kinds of transition-metal complexes that occur in many areas of organometallic chemistry and metalloenzyme chemistry.[110,111] Nevertheless, there are many other important applications where the valence state, magnetic state, or oxidation state of the metal is unknown or changes during a reaction, and the results in Table 13 and the high-B_1 section of Table 14 provide an indication of the ability of various density functionals to treat this important class of problems.

5.13. Self-Exchange. One of the key sources of error in density functional theory is self-exchange.[136−138] For ex-

Table 14. Signed Errors and Mean Unsigned Errors (kcal/mol) in Bond Energies[a]

		$B_1 < 10$					$B_1 > 10$					
methods	X[b]	Cu_2	AgCu	CH_3-H	LiCl	MUE	ZrV	VS	O_3	CN	MUE	MMUE[c]
M05	28	0.0	1.2	−1.7	−2.6	1.4	−12.2	−2.7	−7.2	−1.1	5.8	3.6
B3LYP	20	−5.6	−2.9	−1.7	−4.9	3.8	−19.1	−8.6	−5.8	−1.5	8.8	6.3
B1B95	28	−5.5	−2.6	−0.5	−5.1	3.4	−23.7	−10.1	−5.9	−3.5	10.8	7.1
BLYP	0	−0.8	0.8	−2.8	−5.7	2.5	10.8	4.2	24.0	10.3	12.3	7.4
PBE1PBE	25	−6.3	−3.2	−3.3	−5.4	4.6	−25.1	−11.1	−6.1	−2.0	11.1	7.8
mPWPW91	0	−0.5	1.5	−3.1	−4.7	2.4	9.0	4.9	32.4	12.9	14.8	8.6
mPW1PW91	25	−7.8	−4.5	−3.7	−5.7	5.4	−28.3	−12.7	−11.3	−4.8	14.3	9.8
B1LYP	25	−7.5	−4.9	−3.2	−6.5	5.5	−25.6	−11.4	−19.7	−7.6	16.1	10.8
PBE	0	1.4	3.2	−2.6	−4.3	2.9	13.4	7.0	38.6	16.6	18.9	10.9
M05-2X	56	0.8	3.6	−0.7	1.7	1.7	−42.7	−17.8	−20.6	−7.0	22.0	11.9
BB95	0	2.2	3.7	−0.2	−2.4	2.1	20.3	9.1	41.7	16.1	21.8	12.0
BB1K	42	−8.6	−5.1	−0.6	−5.2	4.9	−41.2	−17.3	−27.3	−12.4	24.6	14.7
average						3.4						9.2

[a] All DFT calculations in this table use the TZQ basis set with consistently optimized geometries. [b] Percentage of Hartree−Fock exchange in each functional. [c] MMUE = 0.5[MUE($B_1 < 10$) + MUE($B_1 > 10$)].

ample, self-exchange is responsible for the poor performance of time-dependent DFT for charge-transfer excited states.[139] One direction of some current research in DFT is to try to obtain correlation functionals that perform well even with 100% Hartree exchange,[38,62,63] which eliminates the self-exchange problem. The present functional does not achieve this goal, but it does perform well with 56% Hartree−Fock exchange, which is much higher than the fraction of Hartree−Fock exchange, 20−28%[7,8,10,15,16,19,22,32,47,56,58,64,65] or, in one case, 31%,[51] of previous functionals with good general-purpose performance and is even higher than the fraction of Hartree−Fock exchange, 42−46%,[27,49−51,58] of functionals designed especially for chemical kinetics. And yet, the M05-2X functional gives better performance than any of the functionals for thermochemical kinetics and alkyl bond energies. Thus, the M05-2X functional should ameliorate some of the problems caused by spurious self-exchange. Furthermore, both M05-2X and M05 are completely free of self-correlation error.

5.14. Comment on Functional Development. A lesson reinforced by the present work is that a good training set is very helpful in parametrizing density functionals, but it is not sufficient. The previous functional forms, prior to M05, did not take full advantage of kinetic energy density and its combination with constraint satisfaction, and they are unable to provide the kind of performance we achieved with M05 and M05-2X. Designing the dependence of the exchange-correlation functional on kinetic energy density was the key to the improved performance achieved here, as compared (for example) to our previous PW6B95 and PWB6K functionals. In designing the new functional form, we built on several key insights in the work of Becke,[10,18,25] but we combined them in new ways and extended them to allow greater flexibility while satisfying the uniform electron gas limit and self-correlation-free limits. In addition, we simultaneously optimized the correlation functional, the exchange functional, and the fraction of Hartree−Fock exchange. It is well-known that it is important for the exchange and correlation functionals to be well-matched. This is partly because they separately have the wrong form at long range and also because exchange density functionals include not only exchange but also near-degeneracy correlation,[7,25,29,140−143] whereas the correlation functional includes only dynamical correlation. It is important to balance the inclusion of near-degeneracy correlation, which is necessary to treat multi-reference character, with Hartree−Fock exchange, which eliminates (or partially eliminates, when $X < 100$ in eq 20) spurious self-exchange interactions.

6. Concluding Remarks

This paper presents a new hybrid meta exchange-correlation functional, M05-2X, for thermochemistry, thermochemical kinetics, and noncovalent interactions. It also presents a more complete picture of the original M05 functional that was originally defined in a preliminary communication. These two functionals incorporate kinetic energy density in a balanced way in the exchange and correlation functionals; they satisfy the uniform electron gas limit, and they are self-correlation-free. They were comparatively assessed against the MGAE109/3 main-group atomization energy database; the IP13/3 ionization potential database; the EA13/3 electron affinity database; the HTBH38/4 database of barrier heights for hydrogen-transfer reactions; the HB6/04 hydrogen-bonding database; the CT7/04 charge-transfer database; the DI6/04 dipole interaction database; the WI7/05 weak interaction database; the PPS5/05 $\pi-\pi$ stacking database; the ABDE4/05 alkyl bond dissociation energy database; the TMAE4/05 database for transition-metal dimers; the MLBE4/05 database for metal−ligand compounds; a dipole moment database, DM6/05, and accurate results for nucleobase interaction energies; the absolute energies of one-electron systems; and the properties of a Lewis acid−base complex, $HCN-BF_3$. From these assessments and from a comparison to results for 28 functionals in the literature, we draw the following conclusions, based on an analysis of mean unsigned errors: (1) The M05-2X, M05, PW6B95, PWB6K, and MPWB1K functionals give the best results for a combination of nonmetallic thermochemical kinetics, thermochemistry, and noncovalent interactions. (2) The M05-2X method gives the best performance for the calculation of absolute and relative bond dissociation energies for single-reference systems and for calculations of noncovalent interactions between nucleobases. (3) The M05 functional gives, in addition, good performance for multireference systems, including metals.

From the present study, we recommend M05-2X, M05, PW6B95, PWB6K, and MPWB1K for general purpose applications in thermochemistry and kinetics, and we espe-

cially recommend M05-2X for calculating bond dissociation energies. For systems involving transition-metal bonding and other multireference systems, we recommend the M05 functional. It is very encouraging that we succeeded in developing density functionals with very broad applicability. They should be especially useful for many applications in chemistry and for condensed-phase systems and molecular recognition problems (including supramolecular chemistry and protein assemblies) where noncovalent interactions are very important.

Acknowledgment. We are grateful to Jan Martin and Mark Iron for assistance with the BMK functional and to Benoit Champagne for supplying the N6 geometry. This work was supported in part by the U.S. Department of Energy, Office of Basic Energy Sciences, by the National Science Foundation, and by the Office of Naval Research.

Note Added after ASAP Publication. This article was released ASAP on February 4, 2006, with an agency that helped support this work not included in the Acknowledgment and with minor errors in Table 13. The correct version was posted on March 7, 2006.

Supporting Information Available: All of the databases are given. The geometries of the nucleobase pairs, of the DM6/05 database, and of the noncovalent interactions are also given. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Becke, A. D. *J. Chem. Phys.* **1986**, *84*, 4524.

(2) Becke, A. D. *Phys. Rev. A: At., Mol., Opt. Phys.* **1988**, *38*, 3098.

(3) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1988**, *37*, 785.

(4) Perdew, J. P. In *Electronic Structure of Solids '91*; Ziesche, P., Eschig, H., Eds.; Akademie Verlag: Berlin, 1991; p 11.

(5) Perdew, J. P.; Wang, Y. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1992**, *45*, 13244.

(6) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 1372.

(7) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648.

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

(9) Gill, P. M. W. *Mol. Phys.* **1996**, *89*, 433.

(10) Becke, A. D. *J. Chem. Phys.* **1996**, *104*, 1040.

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

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

(13) Becke, A. D. *J. Chem. Phys.* **1997**, *107*, 8554.

(14) Adamo, C.; Barone, V. *Chem. Phys. Lett.* **1997**, *274*, 242.

(15) Adamo, C.; Barone, V. *J. Chem. Phys.* **1998**, *108*, 664.

(16) Schmider, H. L.; Becke, A. D. *J. Chem. Phys.* **1998**, *108*, 9624.

(17) Rey, J.; Savin, A. *Int. J. Quantum Chem.* **1998**, *69*, 581.

(18) Becke, A. D. *J. Chem. Phys.* **1998**, *109*, 2092.

(19) Hamprecht, F. A.; Cohen, A. J.; Tozer, D. J.; Handy, N. C. *J. Chem. Phys.* **1998**, *109*, 6264.

(20) Handy, N. C.; Tozer, D. J. *Mol. Phys.* **1998**, *94*, 707.

(21) Krieger, J. B.; Chen, J.; Iafrate, G. J.; Savin, A. In *Electron Correlations and Materials Properties*; Gonis, A., Kioussis, N., Eds.; Plenum: New York, 1999; p 463.

(22) Adamo, C.; Barone, V. *J. Chem. Phys.* **1999**, *110*, 6158.

(23) Perdew, J. P.; Kurth, S.; Zupan, A.; Blaha, P. *Phys. Rev. Lett.* **1999**, *82*, 2544.

(24) Adamo, C.; Ernzerhof, M.; Scuseria, G. E. *J. Chem. Phys.* **2000**, *112*, 2643.

(25) Becke, A. D. *J. Chem. Phys.* **2000**, *112*, 4020.

(26) Rabuck, A. D.; Scuseria, G. E. *Theor. Chem. Acc.* **2000**, *104*, 439.

(27) Lynch, B. J.; Fast, P. L.; Harris, M.; Truhlar, D. G. *J. Phys. Chem. A* **2000**, *104*, 4811.

(28) Proynov, E.; Chermette, H.; Salahub, D. R. *J. Chem. Phys.* **2000**, *113*, 10013.

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

(30) Perdew, J. P.; Schmidt, K. In *Density Functional Theory and Its Applications to Materials*; Van-Doren, V., Alsenoy, C. V., Geerlings., P., Eds.; American Institute of Physics: New York, 2001.

(31) Lynch, B. J.; Truhlar, D. G. *J. Phys. Chem. A* **2001**, *105*, 2936.

(32) Wilson, P. J.; Bradley, T. J.; Tozer, D. J. *J. Chem. Phys.* **2001**, *115*, 9233.

(33) Parthiban, S.; de Oliveira, G.; Martin, J. M. L. *J. Phys. Chem. A* **2001**, *105*, 895.

(34) Boese, A. D.; Handy, N. C. *J. Chem. Phys.* **2002**, *116*, 9559.

(35) Baker, J.; Pulay, P. *J. Chem. Phys.* **2002**, *117*, 1441.

(36) Toulouse, J.; Savin, A.; Adamo, C. *J. Chem. Phys.* **2002**, *117*, 10465.

(37) Boese, A. D.; Martin, J. M. L.; Handy, N. C. *J. Chem. Phys.* **2003**, *119*, 3005.

(38) Becke, A. D. *J. Chem. Phys.* **2003**, *119*, 2972.

(39) Boese, A. D.; Chandra, A.; Martin, J. M. L.; Marx, D. *J. Chem. Phys.* **2003**, *119*, 5965.

(40) Guner, V.; Khuong, K. S.; Leach, A. G.; Lee, P. S.; Bartberger, M. D.; Houk, K. N. *J. Phys. Chem. A* **2003**, *107*, 11445.

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

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

(43) Li, Q. S.; Xu, X. D.; Zhang, S. *Chem. Phys. Lett.* **2004**, *384*, 20.

(44) Staroverov, V. N.; Scuseria, G. E.; Tao, J.; Perdew, J. P. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2004**, *69*, 75102.

(45) Perdew, J. P.; Tao, J.; Staroverov, V. N.; Scuseria, G. E. *J. Chem. Phys.* **2004**, *120*, 6898.

(46) Zhao, Y.; Pu, J.; Lynch, B. J.; Truhlar, D. G. *Phys. Chem. Chem. Phys.* **2004**, *6*, 673.

(47) Xu, X.; Goddard, W. A. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 2673.

(48) Coote, M. L. *J. Phys. Chem.* **2004**, *108*, 3865.

(49) Zhao, Y.; Lynch, B. J.; Truhlar, D. G. *J. Phys. Chem. A* **2004**, *108*, 2715.

(50) Boese, A. D.; Martin, J. M. L. *J. Chem. Phys.* **2004**, *121*, 3405.

(51) Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2004**, *108*, 6908.

(52) Andersson, S.; Gruning, M. *J. Phys. Chem. A* **2004**, *108*, 7621.

(53) Johnson, E. R.; Wolkow, R. A.; DiLabio, G. A. *Chem. Phys. Lett.* **2004**, *394*, 334.

(54) Zhao, Y.; Lynch, B. J.; Truhlar, D. G. *Phys. Chem. Chem. Phys.* **2005**, *7*, 43.

(55) Zhao, Y.; González-García, N.; Truhlar, D. G. *J. Phys. Chem. A* **2005**, *109*, 2012.

(56) Zhao, Y.; Truhlar, D. G. *J. Chem. Theory Comput.* **2005**, *1*, 415.

(57) Schultz, N.; Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2005**, *109*, 4388.

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

(59) Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2005**, *109*, 6624.

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

(61) Perdew, J. P.; Ruzsinszky, A.; Tao, J.; Staroverov, V. N.; Scuseria, G. E.; Csonka, G. I. *J. Chem. Phys.* **2005**, *123*, 62201.

(62) Becke, A. D. *J. Chem. Phys.* **2005**, *122*, 64101.

(63) Dickson, R. M.; Becke, A. D. *J. Chem. Phys.* **2005**, *123*, 111101.

(64) Keal, T. W.; Tozer, D. J. *J. Chem. Phys.* **2005**, *123*, 121103.

(65) Zhao, Y.; Schultz, N. E.; Truhlar, D. G. *J. Chem. Phys.* **2005**, *123*, 161103. Note that, in this communication, we interchanged $c_{C\alpha\beta,i}$ and $c_{C\sigma\sigma,i}$ in Table 1. In addition, "reduced density x_σ" before eq 1 should read "reduced density gradient x_σ".

(66) Schultz, N.; Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2005**, *109*, 11127.

(67) Quintal, M. M.; Karton, A.; Iron, M. A.; Boese, A. D.; Martin, J. M. L. *J. Phys. Chem. A* [ASAP article] **2006**, *110* (2), 709−716.

(68) Lynch, B. J.; Truhlar, D. G. *J. Phys. Chem. A* **2003**, *107*, 8996.

(69) Chakravorty, S. J.; Gwaltney, S. R.; Davidson, E. R.; Parpia, F. A.; Fischer, C. F. F. *Phys. Rev. A: At., Mol., Opt. Phys.* **1993**, *47*, 3649.

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

(71) Izgorodina, E. I.; Coote, M. L.; Radom, L. *J. Phys. Chem. A* **2005**, *109*, 7558.

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

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

(74) Scott, A. P.; Radom, L. *J. Phys. Chem.* **1996**, *100*, 16502.

(75) Champagne, B.; Perpete, E. A.; Jacquemin, D.; Gisbergen, S. J. A. V.; Baerends, E.-J.; Soubra-Ghaoui, C.; Robins, K. A.; Kirtman, B. *J. Phys. Chem. A* **2000**, *104*, 4755.

(76) Bulat, F. A.; Toro-Labbé, A.; Champagne, B.; Kirtman, B.; Yang, W. *J. Chem. Phys.* **2005**, *123*, 14319.

(77) Langhoff, S. R.; Bauschlicher, C. W. *Annu. Rev. Phys. Chem.* **1988**, *39*, 181.

(78) Widmark, O.; Malmqvist, P. A.; Roos, B. *Theor. Chim. Acta* **1990**, *77*, 291.

(79) Pou-Amerigo, R.; Merchan, M.; Nebot-Gil, I.; Widmark, P. O.; Roos, B. *Theor. Chim. Acta* **1995**, *92*, 149.

(80) Pople, J. A.; Head-Gordon, M.; Raghavachari, K. *J. Chem. Phys.* **1987**, *87*, 5968.

(81) Fast, P. L.; Sanchez, M. L.; Truhlar, D. G. *Chem. Phys. Lett.* **1999**, *306*, 407.

(82) Curtiss, L. A.; Redfern, P. C.; Raghavachari, K.; Rassolov, V.; Pople, J. A. *J. Chem. Phys.* **1999**, *110*, 4703.

(83) Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Rassolov, V.; Pople, J. A. *J. Chem. Phys.* **1998**, *109*, 7764.

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

(85) Hehre, W. J.; Radom, L.; Schleyer, P. v. R.; Pople, J. A. *Ab Initio Molecular Orbital Theory*; Wiley: New York, 1986.

(86) Fast, P. L.; Truhlar, D. G. *J. Phys. Chem. A* **2000**, *104*, 6111.

(87) Sinnokrot, M. O.; Sherrill, C. D. *J. Phys. Chem. A* **2004**, *108*, 10200.

(88) Stevens, W.; Basch, H.; Krauss, J. *J. Chem. Phys.* **1984**, *81*, 6026.

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

(90) Schwenke, D. W.; Truhlar, D. G. *J. Chem. Phys.* **1985**, *82*, 2418.

(91) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; 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, 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.01; Gaussian, Inc.: Pittsburgh, PA, 2003.

(92) Werner, H.-J.; Knowles, P. J.; Amos, R. D.; Bernhardsson, A.; Berning, A.; Celani, P.; Cooper, D. L.; Deegan, M. J. O.; Dobbyn, A. J.; Eckert, F.; Hampel, C.; Hetzer, G.; Korona, T.; Lindh, R.; Lloyd, A. W.; McNicholas, S. J.;

Manby, F. R.; Meyer, W.; Mura, M. E.; Nicklass, A.; Palmieri, P.; Pitzer, R.; Rauhut, G.; Schütz, M.; Schumann, U.; Stoll, H.; Stone, A. J.; Tarroni, R.; Thorsteinsson, T. *MOLPRO*, 2002.6; University of Birmingham: Birmingham, AL, 2002.

(93) Stoll, H.; Pavkidou, C. M. E.; Preuss, H. *Theor. Chim. Acta* **1978**, *49*, 143.

(94) Gori-Giorgi, P.; Sacchetti, F.; Bachelet, G. G. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2000**, *61*, 7353.

(95) Gori-Giorgi, P.; Perdew, J. P. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2004**, *69*, 041103.

(96) von-Weizsäcker, C. F. *Z. Phys.* **1935**, *96*, 431.

(97) Yang, G.; Reinstein, L. E.; Pai, S.; Xu, Z.; Carroll, D. L. *Med. Phys.* **1998**, *25*, 2308.

(98) Levy, M.; Perdew, J. P. *Phys. Rev. A: At., Mol., Opt. Phys.* **1985**, *32*, 2010.

(99) Lieb, E. H.; Oxford, S. *Int. J. Quantum Chem.* **1981**, *19*, 427.

(100) Dybala-Defratyka, A.; Paneth, P.; Pu, J.; Truhlar, D. G. *J. Phys. Chem. A* **2004**, *108*, 2475.

(101) Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Pople, J. A. *J. Chem. Phys.* **2000**, *112*, 7374.

(102) Henry, D. J.; Sullivan, M. B.; Radom, L. *J. Phys. Chem. A* **2003**, *118*, 4849.

(103) Hautman, J.; Klein, M. L. *NATO ASI Ser., Ser. E* **1991**, *205*, 395.

(104) Karlin, K. D. *Science* **1993**, *261*, 701.

(105) Crabtree, R. H. *The Organometallic Chemistry of the Transition Metals*, 2nd ed.; Wiley: New York, 1994.

(106) George, S. M. *Chem. Rev.* **1995**, *95*, 475.

(107) Somorjai, G. A. *Chem. Rev.* **1995**, *96*, 1223.

(108) Ratner, M. A.; Davis, B.; Kemp, M.; Mujica, V.; Roitberg, A.; YalirakiI, S. *Ann. N. Y. Acad. Sci.* **1998**, *852*, 22.

(109) *Transition State Modeling for Catalysis*; Truhlar, D. G., Morokuma, K., Eds.; ACS Symposium Series 721; American Chemical Society: Washington, DC, 1999.

(110) Davidson, E. R. *Chem. Rev.* **2000**, *100*, 351.

(111) Siegbahn, P. E. M.; Blomberg, M. R. A. *Chem. Rev.* **2000**, *100*, 421.

(112) Gladysz, J. A. *Chem. Rev.* **2000**, *100*, 1167.

(113) *Handbook on Metalloproteins*; Bertini, I.; Sigel, A.; Sigel, H., Eds.; Dekker: New York, 2001.

(114) Rappe, A. K.; Skiff, W. M.; Casewit, C. J. *Chem. Rev.* **2000**, *100*, 1435.

(115) Coperat, C.; Chabonas, M.; Saint-Arromon, R. P.; Basset, J.-M. *Angew. Chem., Int. Ed.* **2003**, *42*, 156.

(116) Phillips, J. A.; Cramer, C. J. *J. Chem. Theory Comput.* **2005**, *1*, 827.

(117) Reeve, S. W.; Burns, W. A.; Lovas, F. J.; Suenram, R. D.; Leopold, K. R. *J. Phys. Chem. A* **1993**, *97*, 10630.

(118) Fiacco, D. L.; Mo, Y.; Hunt, S. W.; Ott, M. E.; Roberts, A.; Leopold, K. R. *J. Phys. Chem. A* **2001**, *105*, 484.

(119) Hartree, D. R.; Hartree, W.; Swirles, B. *Philos. Trans. R. Soc. London, Ser. A* **1939**, *238*, 229.

(120) Anderson, W. B.; Burdett, J. K.; Czech, P. T. *J. Am. Chem. Soc.* **1994**, *116*, 8808.

(121) McKoy, V.; Sinanoglu, V. *J. Chem. Phys.* **1964**, *41*, 2689.

(122) Laidig, W. D.; Schaefer, H. F. *J. Chem. Phys.* **1981**, *74*, 3411.

(123) Lee, T. J.; Taylor, P. R. *Int. J. Quantum Chem., Quantum Chem. Symp.* **1989**, *23*, 199.

(124) Das, G.; Wahl, A. C. *J. Chem. Phys.* **1967**, *47*, 2934.

(125) Botch, B. H.; Dunning, T. H.; Harrison, J. F. *J. Chem. Phys.* **1981**, *47*, 2934.

(126) Walch, S. P.; Bauschlicher, C. W. *Chem. Phys. Lett.* **1982**, *86*, 66.

(127) Chong, D. P.; Langhoff, S. R. *J. Chem. Phys.* **1986**, *84*, 5606.

(128) Carter, E. A.; Goddard, W. A. *J. Chem. Phys.* **1988**, *88*, 1752.

(129) Murthy, R. B.; Messemer, R. P. *J. Chem. Phys.* **1992**, *97*, 4974.

(130) Schmidt, M. W.; Gordon, M. S. *Annu. Rev. Phys. Chem.* **1998**, *49*, 233.

(131) Staroverov, V. N.; Davidson, E. R. *J. Am. Chem. Soc.* **2000**, *122*, 186.

(132) Gräfenstein, J.; Cremer, D. *Chem. Phys. Lett.* **2000**, *316*, 569.

(133) Pollet, R.; Savin, A.; Leininger, T.; Stoll, H. *J. Chem. Phys.* **2002**, *116*, 1250.

(134) Dunning, T. H.; Harding, L. B. In *Theory of Chemical Reaction Dynamics*; Baer, M., Ed.; CRC press: Boca Raton, FL, 1985; p 1.

(135) Brown, F. B.; Steckler, R.; Schwenke, D. W.; Truhlar, D. G.; Garrett, B. C. *J. Chem. Phys.* **1985**, *82*, 188.

(136) Perdew, J. P.; Zunger, A. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1981**, *23*, 5048.

(137) Kryachko, E. S.; Ludeña, E. V. Energy Density Functional Theory of Many-Electron Systems; Kluwer Academic: Dordrecht, The Netherlands, 1990; p 637.

(138) Ciofini, I.; Adamo, C.; Chermette, H. *J. Chem. Phys.* **2005**, *123*, 121102.

(139) Dreuw, A.; Head-Gordon, M. *Chem. Rev.* **2005**, *105*, 4009.

(140) Slater, J. C. *Phys. Rev.* **1954**, 528.

(141) Tschinke, V.; Ziegler, T. A. *J. Chem. Phys.* **1990**, *93*, 8051.

(142) Ziegler, T. *Chem. Rev.* **1991**, *91*, 651.

(143) Gritsenko, O. V.; Schipper, P. R. T.; Baerends, E. J. *J. Chem. Phys.* **1997**, *107*, 5007.

(144) Slater, J. C. *Quantum Theory of Molecules and Solids. Vol. 4: The Self-Consistent Field for Molecules and Solids*; McGraw-Hill: New York, 1974.

CT0502763

JCTC Journal of Chemical Theory and Computation

Semiempirical Comparative Binding Energy Analysis (SE-COMBINE) of a Series of Trypsin Inhibitors

Martin B. Peters and Kenneth M. Merz, Jr.*

Department of Chemistry, 104 Chemistry Building, The Pennsylvania State University, University Park, Pennsylvania 16802

Received November 21, 2005

Abstract: A scheme to decompose the intermolecular interaction energy of a series of complexes at the semiempirical (SE) level has been developed and validated. The comparative binding energy analysis (COMBINE) (Ortiz, A. R.; Pisabarro, M. T.; Gago, F.; Wade, R. C. *J. Med. Chem.* **1995**, *38*, 2681−2691) and the semiempirical quantum mechanical method pairwise energy decomposition (PWD) (Raha, K.; van der Vaart, A. J.; Riley, K. E.; Peters, M. B.; Westerhoff, L. M. Kim, H.; Merz, K. M., Jr. *J. Am. Chem. Soc.* **2005**, *127*, 6583−6594) were coupled together to form SE-COMBINE. This approach calculates the residue pairwise electrostatic interaction energies, and QSAR models were built with the energies as descriptors using partial least squares (PLS). The application of SE-COMBINE was used as an investigation of the intermolecular interactions between 88 benzamidine inhibitors and trypsin and to test the ability of this new method to predict binding free energies. The predictive capability of SE-COMBINE is shown to be comparable to those of other QSAR methods, and using graphical intermolecular interaction maps (IMMs) enhances the interpretability of receptor-based QSARs.

Introduction

Prediction of the binding free energy of a ligand to a receptor is an unsolved problem. The answer to this problem is to develop a fundamental understanding of receptor−ligand interactions. The accurate prediction of binding free energies requires an exact energy function and a reliable conformational search method that can find the correct binding mode.[1] Considerable research has been carried out in these areas; however, the optimum compromise between computational efficiency and accuracy has yet to be reached.

Computational medicinal chemistry has taken a two-prong approach in the development of new drugs. First, virtual screening procedures, such as the computer-aided structure-based design (CASD) and simple counting methods, are used to screen virtual libraries of 10^6-10^9 molecules. CASD uses docking and scoring to predict the binding mode and affinity of new compounds. Docking methods have been shown to reproduce the binding modes within 2 Å of the crystal structure of protein−ligand complexes.[1] The CASD approach relies on the speed of the scoring function to rapidly evaluate each pose that is generated by docking. The second approach taken by computational chemistry is lead optimization. These methods are routinely carried out using Quantitative Structure Activity Relationship[2] (QSAR) approaches. Most QSAR methods are not receptor-based methods; in other words the receptor is not accounted for in model building. Indeed, this may be the only option if the receptor structure is unknown.

The widely used Comparative Molecular Field Analysis[3] (CoMFA) approach is a grid-based method where molecular properties such as steric (Lennard-Jones) and electrostatic (Coulomb) interactions are calculated between a probe atom and each molecule in the data set at every grid point. The properties at each grid point become descriptors, and models are built using multivariate techniques.

Receptor-based QSAR methods include COMparative BINding Energy analysis[4,5] (COMBINE) and MM/PBSA.[6,7] COMBINE uses a Molecular Mechanics (MM) potential energy function to calculate the intermolecular interactions between the receptor and ligand and builds QSAR models using multivariate statistical tools such as partial least squares (PLS).[8,9]

* Corresponding author e-mail: merz@qtp.ufl.edu. Present address: Department of Chemistry, Quantum Theory Project, University of Florida, 2328 New Physics Building, P.O. Box 118435, Gainesville, FL 32611-8435.

10.1021/ct050284j CCC: $33.50
Published on Web 02/14/2006

The most accurate intermolecular interactions can be obtained using quantum mechanics (QM) methods. High-level QM methods such as Hartree−Fock (HF) and Density Functional Theory (DFT) are frequently used to study small organic systems and protein active sites; however, their use to study protein−ligand interactions is limited due to the high computational cost. Semiempirical (SE) QM methods were developed in the 1970s to reduce the computational cost with minimum loss in accuracy.[10] The most popular SE methods used today are based on the neglect of diatomic differential overlap (NDDO) approximation. The NDDO approximation reduces the number of integral evaluations in QM and in doing so changes the bottleneck of such methods to matrix diagonalization. The divide-and-conquer (D&C),[11−14] density matrix minimization,[15] and localized molecular orbital[16] methods have been developed to address the problem of matrix diagonalization enabling the application of SE methods to macromolecular systems. The D&C approach has been implemented in the program DivCon[17] which uses the SE Hamiltonians AM1,[18] PM3,[19,20] MNDO/d,[21,22] and PM3-PDDG.[23] Recently, D&C methods such as QMSCORE,[24,25] pairwise energy decomposition (PWD),[26] and DCNMR[27] have been developed to study protein−ligand interactions in DivCon. QMSCORE is a SE based score function that outperforms other score functions such as AutoDock and DrugScore. The PWD method is a novel approach where the electrostatic interaction energy is partitioned into self- and cross components between atoms. PWD has successfully been used to investigate the effect of binding of a series of fluorine-substituted ligands to human carbonic anhydrase II. DCNMR has been shown to predict NMR chemical shifts from the 3D structure of protein−ligand complexes.

In this work the PWD method was coupled to the COMBINE method creating SE-COMBINE to study a large set of protein−ligand complexes at the SE level of theory. PWD calculates the pairwise electrostatic interactions between a protein and ligand using the linear-scaling D&C approach. Similar to the COMBINE method, the residue pairwise energies were used to build QSAR models. The utility of SE-COMBINE was demonstrated by investigating the structure−activity relationship of a series of trypsin-like serine protease inhibitors.

Serine proteases are involved in many processes in the body such as protein digestion and blood coagulation.[28] The serine protease family of enzymes catalyzes protein hydrolysis. Trypsin, chymotrypsin, and elastase are common enzymes involved in the digestion of dietary proteins. Thrombin, factor Xa, and plasmin are key enzymes of the blood-clotting cascade. They differ only by their selectivity; for example, trypsin regiospecifically hydrolyzes at the carboxyl side of lysine and arginine amino acids, whereas chymotrypsin cleaves at aromatic sites. All serine proteases contain the catalytic triad Asp102, His57, and Ser195, which allows the catalytic cleavage of peptide bonds through an acyl intermediate as shown in Figure 1. The neighboring aspartic acid and histidine residues modify the serine from a hydroxyl to an alkoxide allowing the nucleophilic attack of the carbonyl group to occur.[29] The development of inhibitors of trypsin-like serine proteases has been an active area of research because they are important targets in the blood-clotting cascade and also serve as a useful model system to study protein−ligand interaction.

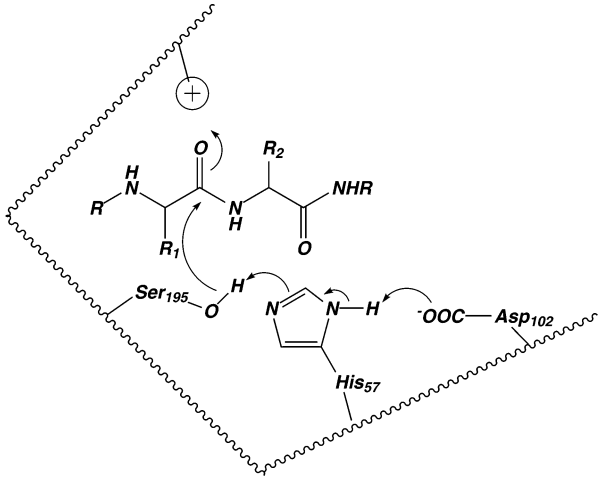


Figure 1. Trypsin hydrolysis mechanism. The general acid/base catalysis takes place using the catalytic triad of Asp102, His57, and Ser195. Asp102 removes a proton from His57, which activates Ser195 from a hydroxyl to an alkoxide nucleophile [adapted from Silerman, 2002].

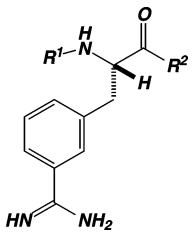


Figure 2. The structure of the 3-amidinophenylalanine molecule. Structural changes occur at two positions, R_1 and R_2 [adapted from Böhm et al.].

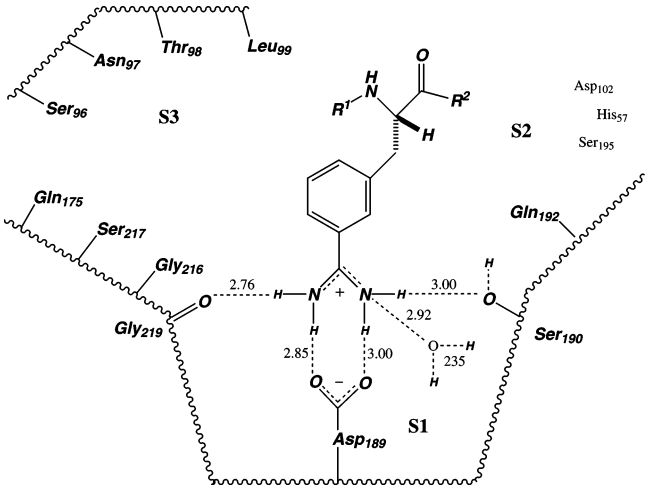
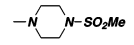
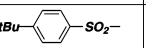
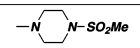
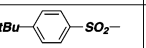
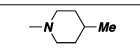
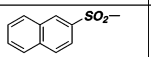
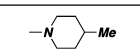
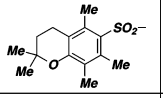
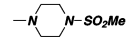
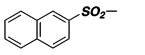
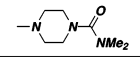
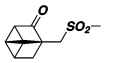
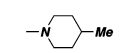
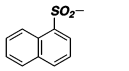
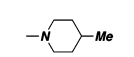
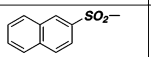
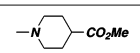
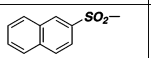
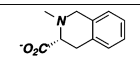
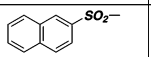
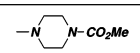
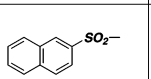
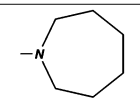
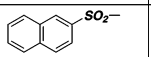
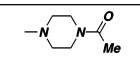
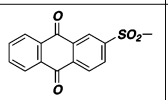
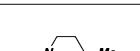
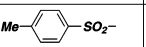
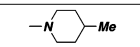
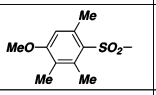
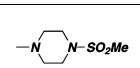
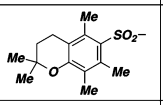
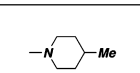
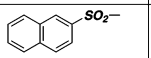
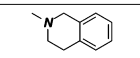
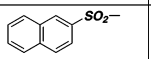
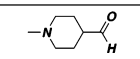
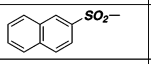
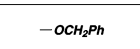


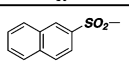
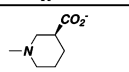
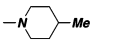
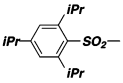
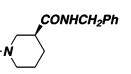
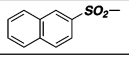
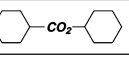
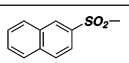
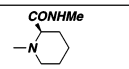
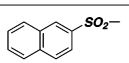
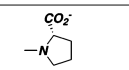
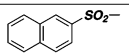
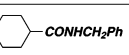
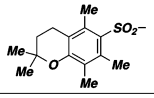
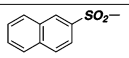
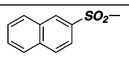
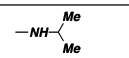
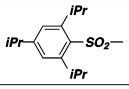
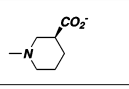
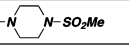
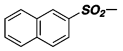
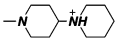
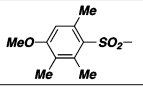
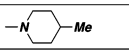
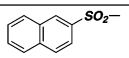
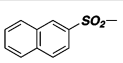
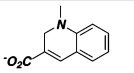
Figure 3. Schematic representation of 3-amidinophenylalanine bound to trypsin. The distances shown are determined from the complex of 3-TAPAP (Brookhaven Protein Data Bank reference: 1PPH) where R^1 is tosyl and R^2 is piperidine. Distances shown are in angstroms [adapted from Böhm et al.].

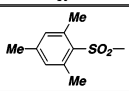
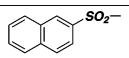
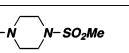
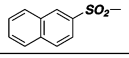
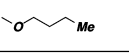
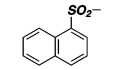
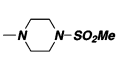
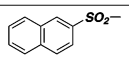
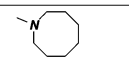
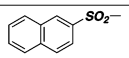
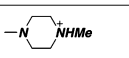
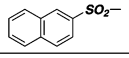
Table 1: 88 Trypsin Inhibitors[a]

No.	R^1	R^2	Charge	pK_i
1			+1	6.770
2			+1	6.796
3			+1	6.699
4			+1	6.854
5			+1	6.119
6			+1	6.770
7			+1	6.201
8			+1	6.201
9			+1	7.444
10			0	6.886
11			+1	7.699
12			+1	6.260
13			+1	6.854
14			+1	7.131
15			+1	6.284
16			+1	5.745
17			+1	6.137
18			+1	6.585
19			+1	6.658
20			+1	6.284
21			+1	6.678
22			+1	5.959
23			+1	5.398
24			+1	6.481
25			+1	6.161
26			+1	6.108

No.	R^1	R^2	Charge	pK_i
27			+1	5.658
28			+1	5.854
29			0	5.347
30			+1	5.824
31			+1	5.398
32			+1	6.409
33			+1	6.569
34			+1	6.796
35			+1	6.004
36			+1	5.921
37			+1	7.174
38			+1	6.215
39			+1	6.102
40			0	6.201
41			+1	5.602
42			+1	6.921
43			+1	5.921
44			+1	5.444
45			+1	5.921
46			+1	5.658
47			+1	5.678
48			+2	6.658
49			+1	6.367
50			0	6.237
51			+1	6.000
52			+1	5.092
53			+1	5.921
54			+1	6.071

Table 1: (Continued)

No.	R^1	R^2	Charge	pK_i
55			0	6.357
56			+1	4.854
57			+1	6.337
58			+1	7.097
59			+1	5.102
60			0	4.796
61			+1	6.569
62		—NHMe	+1	4.495
63		—NHMe	+1	4.602
64			+1	4.796
65			0	5.620
66	H—		+2	4.538
67			+2	6.000
68		—NHMe	+1	3.854
69	H—		+2	4.538
70		$-CO_2^-$	0	3.928
71			0	4.509

No.	R^1	R^2	Charge	pK_i
72		—NHMe	+1	3.000
73			+1	6.721
74			+1	6.585
75			+1	6.495
76			+1	6.215
77			+2	5.886
78			+1	6.357
79			+1	5.721
80			+1	6.149
81			+1	6.509
82			+1	6.009
83			+1	6.796
84			+1	7.569
85			+1	5.745
86			+1	7.638
87		—NHMe	+1	4.585
88		—NHMe	+1	4.337

[a] Substituents at positions at R^1 and R^2, formal charges and pK_i values are listed [adapted from Böhm et al.]

Trypsin is synthesized in the pancreas as a zymogen (inactive enzyme) called trypsinogen. When required, trypsinogen is secreted into the small intestine through the bile duct and after enzymatic removal of an N-terminal amino acid sequence trypsin (24kDa) is formed. Trypsin has a large binding pocket, **S1**, adjacent to the catalytic site with an aspartic acid at the base. This pocket favors the binding of the positively charged amino acids, lysine and arginine. The strong ionic interaction allows for the cleavage reaction to take place. The enzymes thrombin and factor Xa have similar **S1** pockets; however, the **S2** and **S3** pockets vary in composition and in size (thrombin has the insertion loop, Tyr60A-Trp60D, whereas the others do not).[30] Thus the development of selective inhibitors of trypsin, thrombin, and factor Xa has posed a challenge for both experimental and computational research.[31]

Since 1965, benzamidine-based inhibitors of trypsin-like proteases have been developed.[32] The amidinophenylalanine group mimics the guanidinalkyl functional group of arginine as shown in Figure 2. The X-ray structure of *N*α-[4-toluene sulfonyl]-L-*m*-amidino-phenylalanyl (3-TAPAP, a 1.2 μmol/L inhibitor) bound to trypsin (1.9 Å resolution) was reported in 1991 (PDB: 1PPH).[30] A schematic representation of the key interactions between the 3-amidinophenylalanine group and the **S1** pocket of trypsin from 1PPH is shown in Figure 3. The amidino group forms a near symmetric salt bridge with Asp189 and is also hydrogen bonded to both Gly219@O and a water molecule. The phenyl ring is sandwiched between the sequences Ser190-Gln192 and Trp215-Gly216. Gly216 forms hydrogen bonds with the amino and carbonyl group of the *m*-amidino-phenylalanine moiety. The tosyl group, R^1, fills the **S3** pocket and lies perpendicular to the indole group of Trp215, while anoxygen of the sulfonyl group points toward Gly219@N. The piperidine group occupies the **S2** pocket and is flanked on either side by His57 and the toluene group of the tosyl moiety. Using the benzamidine scaffold, a series of inhibitors was reported,[33,34] and recently 3D QSAR techniques such as CoMFA[35], CoMSIA,[36] and QSM[37] have been used to investigate the inhibitor selectivity between thrombin, trypsin, and factor Xa.[38]

Trypsin and its inhibitors are very well characterized, i.e. protein−ligand complex structures and binding affinity data

are available, thus providing an excellent starting point for a computational study. The semiempirical quantum mechanical decomposed intermolecular interactions between trypsin and a series of inhibitors shown in Table 1 were examined in this study. Using the protein-residue−ligand-fragment interaction energies, a comparative binding energy analysis was carried out using PLS to build a receptor-based 3D-QSAR model.

Computational Approach

Consider the interaction of a receptor R, with a ligand L, to form the complex R ·L:

$$\mathrm{R} + \mathrm{L} \rightleftharpoons \mathrm{R{\cdot}L} \tag{1}$$

The interaction energy can be calculated using the following

$$E_{\mathrm{INT}} = E_{\mathrm{R{\cdot}L}} - (E_{\mathrm{R}} + E_{\mathrm{L}}) \tag{2}$$

where $E_{\mathrm{R{\cdot}L}}$ is the energy of the complex, and E_{R} and E_{L} are the energies of the receptor and ligand, respectively. Equation 2 can also be represented with ΔE_{R} and ΔE_{L} as the change in energy of the receptor and ligand upon binding as in eq 3.

$$E_{\mathrm{INT}} = E_{\mathrm{R{\cdot}L}} + \Delta E_{\mathrm{R}} + \Delta E_{\mathrm{L}} \tag{3}$$

Carrying out a residue-based pairwise decomposition of the interaction energy leads to the following

$$E_{\mathrm{INT}} = \sum_I \sum_J E_{IJ} + \sum_I \sum_{K<I} \Delta E_{IK} + \sum_J \sum_{L<J} \Delta E_{JL} + \sum_I \Delta E_I + \sum_J \Delta E_J \tag{4}$$

where I is the index of the residues in the receptor, J is the index of fragments in the ligand, E_{IJ} is the residue(receptor)-fragment(ligand) interaction energy (a true cross term, this term is only present in the complex), ΔE_{IK} is the change in the interresidue energy upon binding of the residues in the receptor, ΔE_{JL} is the change in the interfragment energy upon binding of the fragments in the ligand, ΔE_{I} is the change in intraresidue energy upon binding of the residues in the receptor, and ΔE_{J} is the change in intrafragment energy upon binding of the fragments in the ligand.

Considering the above in terms of a classical molecular mechanics force field, the first term would be the electrostatic and van der Waals interactions between receptor residues and ligand fragments. The second and third terms would be change in electrostatic and van der Waals interactions between receptor residues and other residues and ligand fragments and other fragments upon complexation. The fourth and fifth terms would be the change in the bond, angle, torsion, and nonbonded interactions of receptor residues and ligand fragments.

Within a semiempirical approach the binding energy expression of eq 4 can be expressed in terms of the quantities derived by Raha et al.[26]

$$E_{\mathrm{INT}} = \sum_I \sum_J \Big(\sum_A \sum_B E_{AB} + E'_{AB} + E^{core}_{AB}\Big) + A \in I, B \in J$$

$$\sum_I \sum_{K<I} \Big(\sum_A \sum_B \Delta E_{AB} + \Delta E'_{AB} + \Delta E^{core}_{AB}\Big) + A \in I, B \in K$$

$$\sum_J \sum_{L<J} \Big(\sum_A \sum_B \Delta E_{AB} + \Delta E'_{AB} + \Delta E^{core}_{AB}\Big) + A \in J, B \in L$$

$$\sum_I \Big(\sum_A \{\Delta E_A + \sum_{B<A} \Delta E_{AB} + \Delta E'_{AB} + \Delta E^{core}_{AB}\}\Big) + \mathrm{A,B} \in \mathrm{I}$$

$$\sum_J \Big(\sum_A \{\Delta E_A + \sum_{B<A} \Delta E_{AB} + \Delta E'_{AB} + \Delta E^{core}_{AB}\}\Big)\, A,B \in \mathrm{J} \tag{5}$$

where

$$E_A = \frac{1}{2}\sum_\mu^A \sum_\nu^A P^{AA}_{\mu\nu}\left(2H^{AA}_{\mu\nu} + \sum_{\lambda\sigma}^A P^{AA}_{\lambda\sigma}\left[(\mu^A\nu^A|\sigma^A\lambda^A) - \frac{1}{2}(\mu^A\sigma^A|\lambda^A\nu^A)\right]\right) \tag{6}$$

$$E'_{AB} = \sum_\mu^A \sum_\nu^A \sum_{\lambda\sigma}^B P^{AA}_{\mu\nu} P^{BB}_{\lambda\sigma}(\mu^A\nu^A|\sigma^B\lambda^B) \tag{7}$$

$$E_{AB} = \sum_\mu^A \sum_\nu^B P^{AB}_{\mu\nu}\left(2H^{AB}_{\mu\nu} - \frac{1}{2}\sum_\lambda^B \sum_\sigma^A P^{BA}_{\lambda\sigma}(\mu^A\sigma^A|\lambda^B\nu^B)\right) \tag{8}$$

$$E^{core}_{AB} = \sum_A \sum_{B<A} \frac{Z_A Z_B}{R_{AB}} \tag{9}$$

PWD calculates the self-energy of the atom, E_A, core−electron interactions, E^{core}_{AB}, electron−electron repulsions, E'_{AB}, and exchange between the atoms, E_{AB}, as shown in eqs 6−9. H is the one-electron matrix, F is the Fock matrix, and P is the density matrix. Z is the nuclear charge on an atom, R_{AB} is the atomic separation between A and B. Equation 5 represents the decomposition of the semiempirical interaction energy between a receptor and ligand. The E_A term has a large negative energy contribution to the total energy since it contains the one-center terms. E'_{AB} contains all the electronic repulsion, and so it is a positive contributor to the energy which comes from the diagonal block of the Fock matrix. E_{AB} contains the exchange repulsion between atoms and is a small negative contributor to the total energy, which stems from the off-diagonal elements of the Fock, one-electron, and density matrices. As originally described, it contains most of the binding energy. Accepting, as an approximation, that the receptor and ligand conformations remain the same upon binding, the decomposed energy becomes

$$E_{\mathrm{INT}} = \sum_I \sum_J (E_{AB} + E'_{AB} + E^{core}_{AB}) + \sum_I \sum_{K<I} (\Delta E_{AB} + \Delta E'_{AB}) + \sum_J \sum_{L<J} (\Delta E_{AB} + \Delta E'_{AB}) + \sum_I (\Delta E_A + \Delta E_{AB} + \Delta E'_{AB}) + \sum_J (\Delta E_A + \Delta E_{AB} + \Delta E'_{AB}) \tag{10}$$

Equation 10 as written implies a sum over A and B in the first 3 terms, over A in the fourth term, and over B in the final term.

Mol	Act	IJ-E'_{AB}	IJ-E_{AB}	IJ-E^{core}_{AB}	IK-E'_{AB}	IK-E_{AB}	JL-E'_{AB}	JL-E_{AB}	I-E'_{AB}	I-E'_{AB}	I-E^{core}_{AB}	J-E'_{AB}	J-E_{AB}	J-E^{core}_{AB}
1	•	·	·	·	·	·	·	·	·	·	·	·	·	·
2	•	·	·	·	·	·	·	·	·	·	·	·	·	·
g	•	·	·	·	·	·	·	·	·	·	·	·	·	·
N	•	·	·	·	·	·	·	·	·	·	·	·	·	·

Figure 4. Schematic diagram of an example data table used in SE-COMBINE. The size of the descriptor matrix is defined by the number of compounds, *N*, and by the number of descriptors. The experimental data, Act, is a single column in the data table, while the indices I, J, K, and L refer to eq 10. IJ − E'_{AB}, IJ − E_{AB}, and IJ − E^{core}_{AB} are energy terms between receptor residues and ligand fragments, true cross terms. IK − E'_{AB} and IK − E_{AB}, are energy terms between pairs of receptor residues. JL − E'_{AB} and JL − E_{AB} are energy terms between pairs of ligand fragments. The remaining terms are residue and fragment self-energy terms.

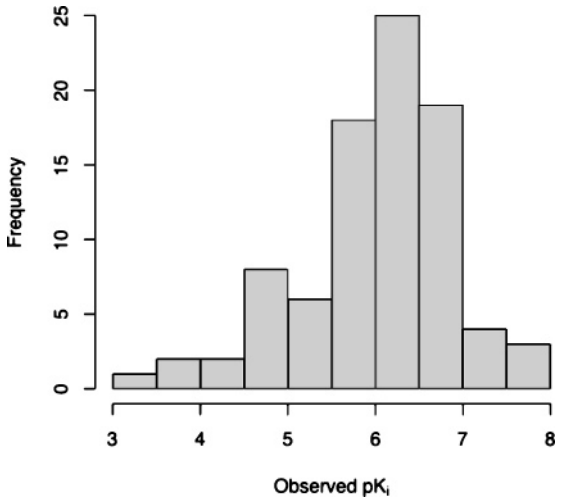


Figure 5. Trypsin inhibitor activity frequency distribution. Affinities spread over a 4.7 logarithm unit's range, which allows a statistically significant 3D QSAR to be derived.

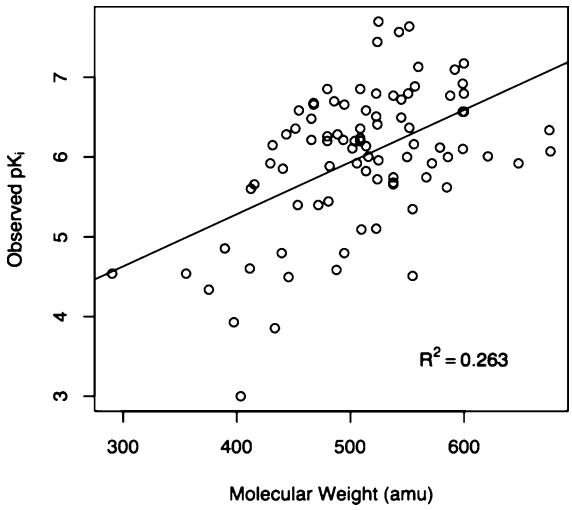


Figure 6. Trypsin inhibitors pK_i versus molecular weight. Poor correlation is observed with an R^2 value of 0.26.

The pwdPy program was developed in order to perform a pairwise decomposition of the interaction energy between the ligand fragments and the protein residues using the formalism described above in eq 10. That is, the program was used to read the DivCon output of the ligand, protein, and complex calculations and to produce a descriptor table similar to the one shown in Figure 4.

Procedure

(1) Data Set. The crystal structure of 3-TAPAP bound to trypsin (1PPH) was obtained from the Protein Data Bank (PDB). 3-TAPAP is a 3-amidinophenylalanine based inhibitor of trypsin. Eighty-eight compounds (coordinates and activity data), including 3-TAPAP (all fully protonated), which bind to trypsin, were kindly provided by Prof. Gerhard Klebe. The L-conformations of the central phenylalanine are more potent by a factor of 50−100 over the D-conformations; however, the pK_i reported are mixtures of the L and D forms.[34] Only the L-conformations of the compounds were used in this study. All 88 structures share a common core and differ at positions R^1 and R^2 as shown in Figure 2. The structures of the R^1 and R^2 groups for all compounds and their activity data are listed in Table 1. The affinities of the inhibitors spread over a range of 4.7 pK_i units. However, as shown in Figure 5 the majority of the inhibitors' affinities lie in the range between 5.5 and 7.0 pK_i units. Note that the variation in size of the R groups does not translate to a molecular weight dependence on binding affinity as shown in Figure 6.

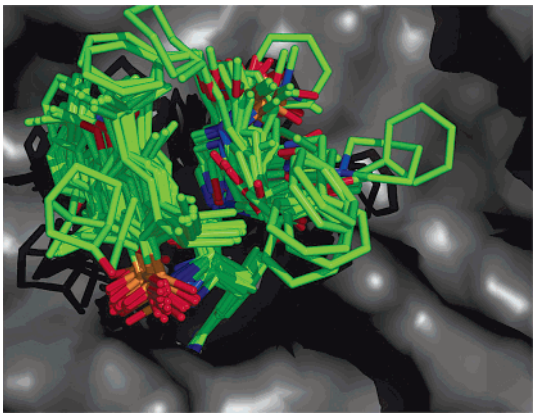


Figure 7. Inhibitors aligned in the active site of trypsin. Trypsin is represented as a surface and inhibitors as sticks. Hydrogen atoms are not shown for clarity [adapted from Böhm et al.].

(2) Molecular Mechanics Modeling of the Receptor. The 3-TAPAP structure and all water molecules except number 235 were removed from the 1PPH crystal structure. All

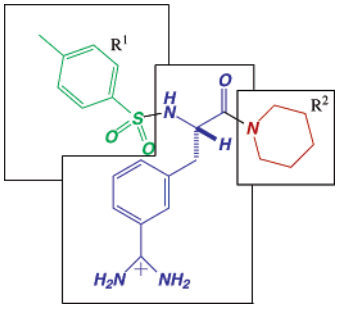


Figure 8. Schematic diagram of a trypsin inhibitor fragmentation. The structure in blue is the 3-amidino-phenylalanine moiety (**APM**). The **TOS** group is colored green, while the **PIP** group is shown in red.

references to the amino acid names and numbers follow that used in 1PPH. Wat235 is the characteristic water molecule present in the **S1** pocket. Hydrogen atoms were added to the protein using the LEaP module of AMBER, followed by a hydrogen minimization (1500 steps) using the SANDER module of AMBER 8.[39] All acidic residues were assumed to be deprotonated while all basic residues were protonated.

(3) Molecular Mechanics Modeling of the Complexes. The 88 ligands were aligned onto 3-TAPAP (**45**) using the 3-amidinophenylalanine moiety as a template. Each ligand was placed in the active site of 1PPH as shown in Figure 7. The 88 compounds vary in size and shape, and so some close contacts were expected. To correct this the inhibitors were allowed to relax in the active site using a restrained minimization of 1500 steps (500 steepest descent followed by 1000 conjugate gradient), followed by a full minimization of all atoms in the system (500 steepest descent followed by 1000 conjugate gradient steps) using AMBER.

(4) Semiempirical D&C Calculations. Semiempirical D&C calculations were performed using the PM3 Hamiltonian within the program DivCon. Five calculations were performed for each of the 88 compounds in the data set: (1) protein only, (2) ligand (1 fragment), (3) ligand (3 fragments), (4) complex (ligand with 1 fragment), and (5) complex (ligand with 3 fragments). The protein was divided into subsystems based on the standard amino acid residue definitions. All atoms of the ligand were grouped into one fragment in calculation 2, and the fragment name was set to **TAP**. Each ligand in the data set was also divided into three groups or fragments, e.g. **45** is shown in Figure 8. The first fragment consists of the 3-amidinophenylalanine moiety (**APM**), the R^1 group contained aryl sulfonyl groups (**TOS**), and the third contains either piperidine or piperazine groups (**PIP**). The fragments were named based on those residues found in 3-TAPAP of 1PPH. A cutoff for the Fock matrix of 20 Å and a divide-and-conquer buffering scheme of 4.2/2 Å were used throughout the entire study. Note that the total

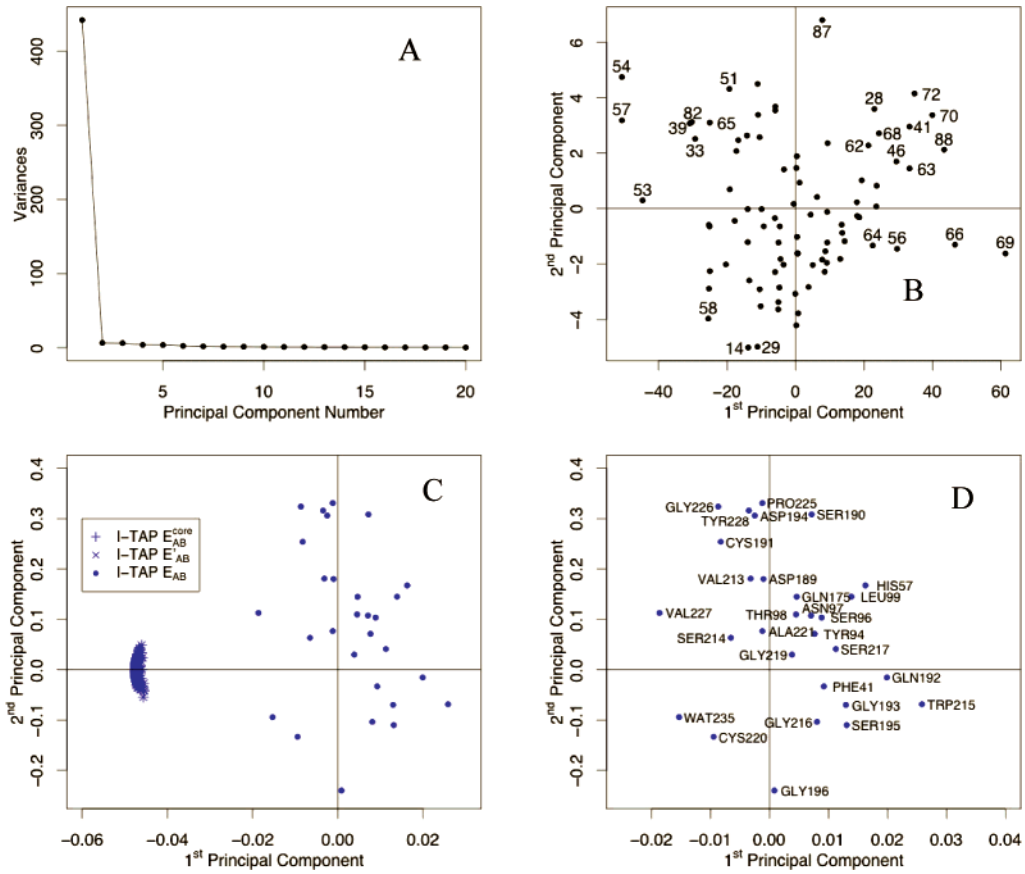


Figure 9. Model Lig1C PCA results. (A) Scree plot. (B) Score plot of PC 1 versus 2. Points representing complexes of interest are labeled using ligand numbers. (C) Loading plot of PC 1 versus 2. The types of descriptors in this model are shown in the legend where I denotes any amino acid in the protein. (D) Loading plot of PC 1 versus 2 where only E_{AB} descriptors are considered. Labels shown are the protein residue name and number involved in the interaction with the ligand.

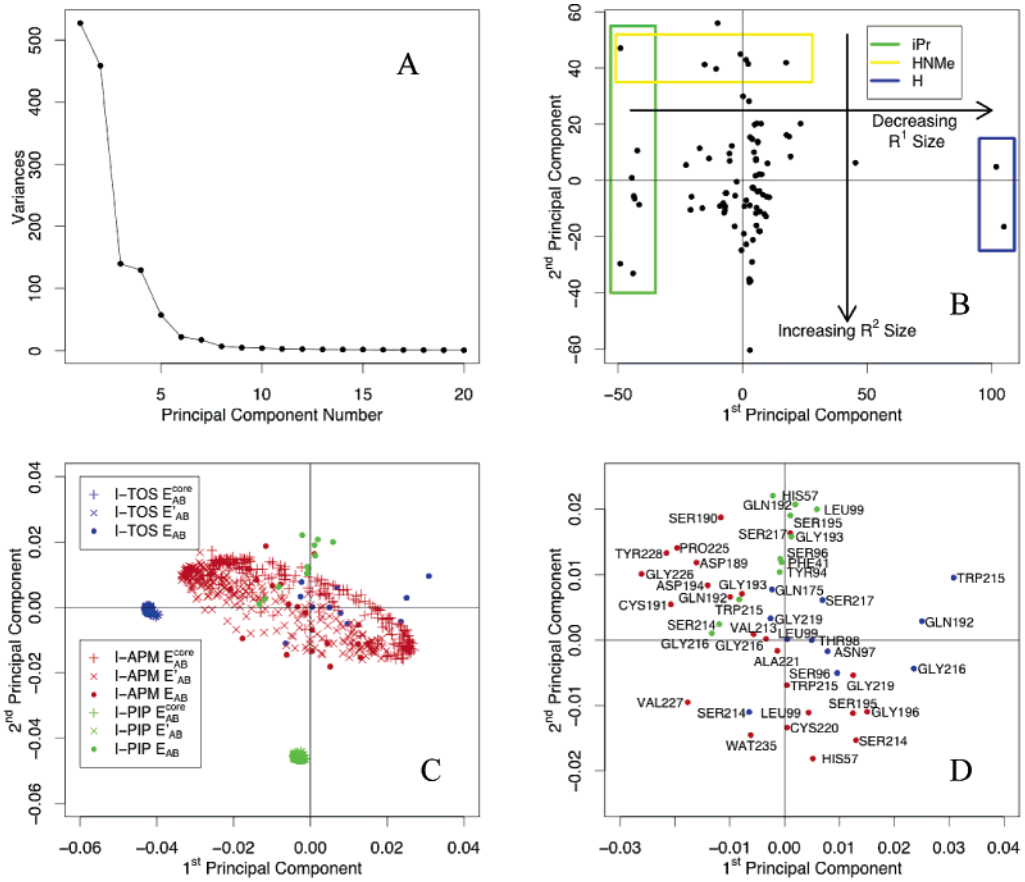


Figure 10. Model Lig3C PCA results. (A) Scree plot. (B) Score plot of PC 1 versus 2. Points representing complexes of interest are labeled using ligand numbers. Highlighted are isopropyl, ^{i}Pr, and hydrogen, H, on R^1 and HNMe on R^2. (C) Loading plot of PC 1 versus 2. The types of descriptors in this model are shown in the legend where I denotes any amino acid in the protein. (D) Loading plot of PC 1 versus 2 where only E_{AB} descriptors are considered. Descriptors are labeled by protein residue, whereas the colors represent the fragments of the ligand.

interaction energy between the receptor and ligand does not change after fragmentation.

(5) Chemometric Analysis. The pwdPy program was used to pairwise-decompose the interaction energy between the ligand fragments and the protein residues. The statistical analysis of the decomposed energies was preformed using the program R.[40]

As a first step, Principal Component Analysis (PCA) was carried out to examine the distribution of the complexes in the descriptor space. The similarity/dissimilarity between inhibitors was investigated using score plots. The descriptor pool was pruned to remove descriptors that returned zero values. Auto-scaling was applied to the descriptor matrices, or, in other words, each descriptor was processed to have a mean of zero and a standard deviation of one. This ensured that certain variables did not dominate due to their magnitude. PLS models were built to explore the structure−activity relationship of the inhibitors. Internal validation was carried out using leave-one-out (LOO) cross-validation, and the optimal dimensionality of each model was assigned from its cross-validated predictive ability. External validations were also carried out where 10 structures were removed randomly from the original data matrix to become the prediction set, while the rest remained as the training set. Ten such prediction sets were generated. Descriptor pruning and auto-scaling was applied to the training set following the above procedure. After the models were generated using the training set, the prediction set was autoscaled using the means and standard deviations from the training set. Another external validation study was carried out where the training and prediction sets were predefined by Böhm et al. This was used in order to compare SE-COMBINE to methods such as CoMFA, CoMSIA, and quantum similarity.

Together with the square of the Pearson's correlation coefficient, R^2, and the cross-validated correlation coefficient, Q^2, the standard deviation of error of calculations, SDEC, and the standard deviation of error prediction, SDEP, were used to assess the quality of the models. SDEP can also be defined as the root-mean-squared error of the dependent variables in a LOO scheme or external data set. Similarly, SDEC is calculated for those variables used to build the model or training set. For each model, the biological activities of the inhibitors were scrambled randomly, and the activities were predicted, as a way of detecting the possibility of chance correlation. And in all cases only negative Q^2 values were observed.

Results

The pairwise interactions between the 224 amino acid residues of trypsin and a water molecule with each inhibitor

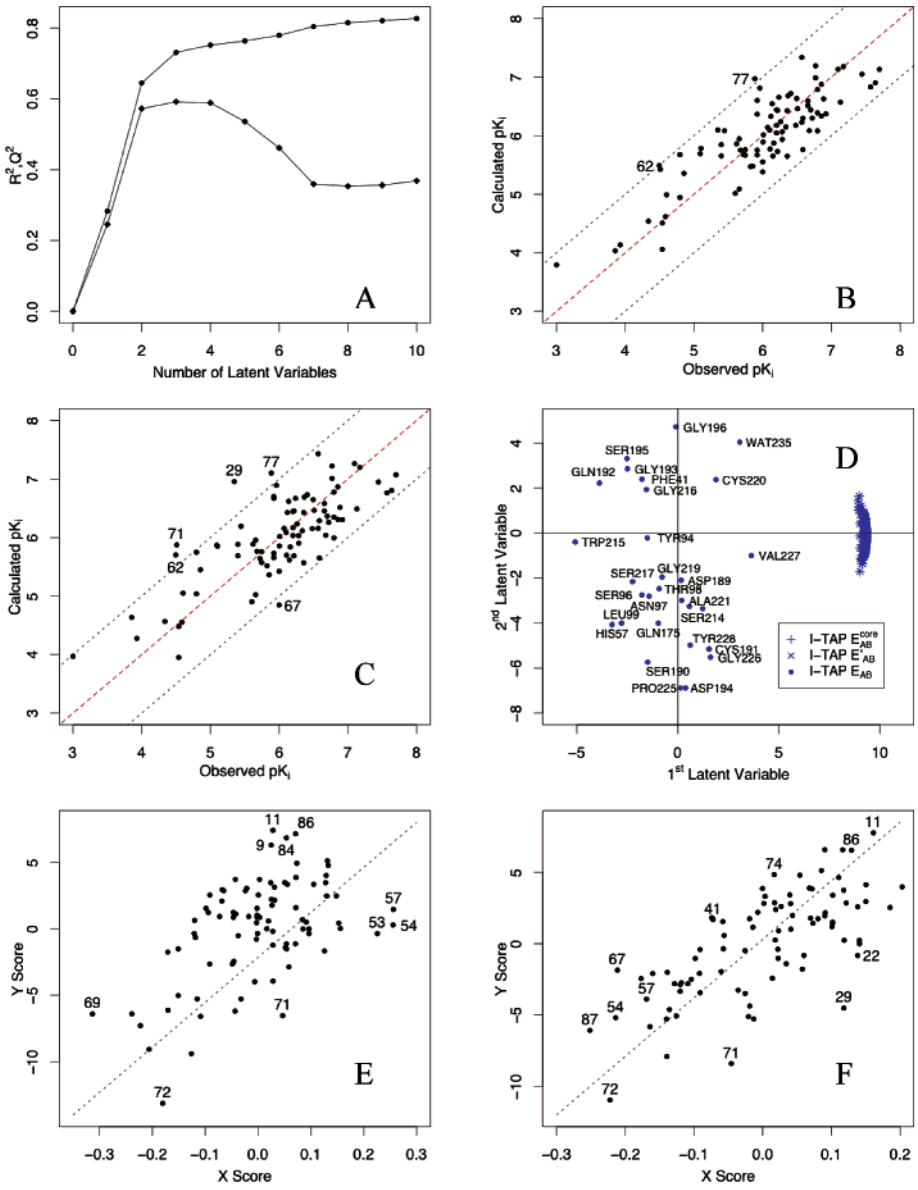


Figure 11. Model Lig1C PLS results. (A) R^2 and Q^2 versus number of latent variables. R^2 is represented as solid circles and Q^2 as solid diamonds. (B) Observed versus calculated pK_i values from internal validation. The red line represents the optimal correlation, while the blue lines are a pK_i unit from the optimal line. (C) Observed versus calculated pK_i values from LOO cross-validation. (D) Loading plot of the first and second latent variables. (E) X-score versus Y-score for the first latent variable. (F) X-score versus Y-score for the second latent variable.

were calculated. The effect of fragmentation of the ligand structure was investigated by considering a single (Lig1) and triple fragment (Lig3) scheme. The Lig1 scheme yields a total of 25 878 (model Lig1A) descriptors (computed using eq 11) to fully decompose E_{INT} (eq 10). Only considering the cross term E_{IJ}, this number reduces to 672 (model Lig1B). It was found that the majority of the E_{AB} terms were zero [Tests with a water dimer showed that E_{AB} is zero with oxygen−oxygen distances greater than 4 Å.]; therefore, a E_{AB} descriptor was removed if more than 95% of its terms were zero. The remaining descriptors were auto-scaled because the E_{AB}, E'_{AB}, and E_{AB}^{core} terms span different ranges (model Lig1C). Using the same procedure the Lig3 scheme produces 52 655 descriptors (model Lig3A). This number was reduced to 2016 when only E_{IJ} interactions were considered (model Lig3B). The E_{AB} terms were pruned reducing the dimensionality, and autoscaling was applied, thus producing model Lig3C. The total number of the descriptors per model is given in Table 2.

Table 2: Number of Descriptors per QM-COMBINE Model

	number of descriptors	
model	Lig1	Lig3
A	25878	52655
B	672	2016
C	477	1389

$$((I*J) + I + J)*3 + \sum_{n}^{I}(n-1)*2 + \sum_{n}^{J}(n-1)*2 \qquad (11)$$

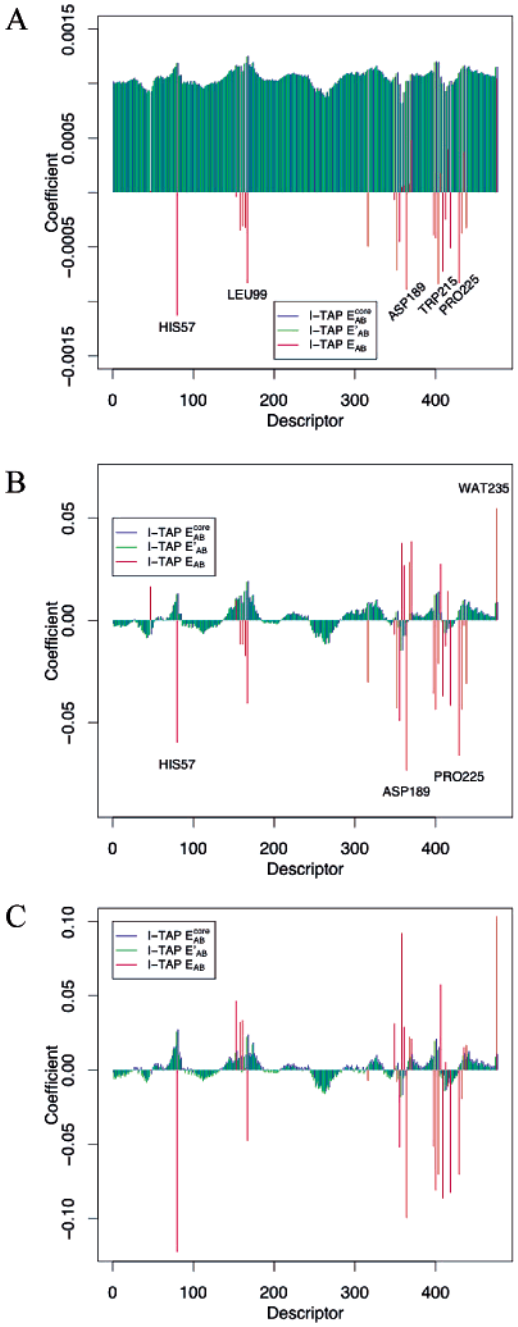


Figure 12. Model Lig1C PLS coefficient plots. (A) Latent variable 1. (B) Latent variable 2. (C) Latent variable 3.

Model Lig3A contained 52 655 descriptors. This model could not be handled using current computer hardware (required over 4 GB of RAM) and was skipped. The same holds for model Lig1. Models Lig1B and Lig3B were statistically compromised since the majority of the E_{AB} terms were zero. Therefore the first model to be considered was Lig1C and then Lig3C.

Principal Component Analysis: Model Lig1C. A PCA of the descriptor matrix for the 88 complexes was performed using the autoscaled variables.[41,42] The scree plot of the PCA shown in Figure 9.A illustrates that two principal components (PCs) successfully models the data. PC 1 and 2 explain 94.2% of the variance in the *X* matrix. The score plot of the first and second PC is shown in Figure 9.B. The score plot is divided into quadrants; the upper right quadrant contains complexes that have positive scores for both PC 1 and PC 2. These complexes all have small R^2 groups, while the lower left quadrant has large R^2 groups. A similar trend is observed for the R^1 groups: the top left quadrant contains large R^1 groups, while the lower right contains complexes with smaller R^1 groups. The complexes are distinguished in the descriptor space, with R group size playing a relevant role.

Table 3: Models Lig1C and Lig3C PLS Results[a]

model	LV	*X* variance (cumulative)	R^2	Q^2	SDEC	SDEP	$SDEC^{ext}$	$SDEP^{ext}$
Lig1C	1	92.88	0.28	0.25	0.74	0.76	0.75	0.68
	2	94.18	0.65	0.57	0.57	0.57	0.46	0.47
	3	94.68	0.73	0.59	0.45	0.56	0.32	0.42
Lig3C	1	32.53	0.40	0.32	0.68	0.72	0.69	0.60
	2	45.21	0.50	0.42	0.62	0.67	0.62	0.58
	3	79.64	0.51	0.43	0.61	0.66	0.61	0.58
	4	82.86	0.63	0.48	0.53	0.64	0.51	0.59
	5	89.63	0.67	0.53	0.50	0.60	0.46	0.58
	6	94.76	0.70	0.55	0.48	0.59	0.42	0.55
	7	96.38	0.74	0.58	0.45	0.57	0.38	0.51

[a] LV represents the number of latent variables in the model. The optimal number of LVs is shown in bold. R^2 and Q^2 represent the correlation coefficient of training and LOO. SDEC and SDEP are the standard deviations of calculation and prediction for internal validation. $SDEC^{ext}$ and $SDEP^{ext}$ are similarly defined applied to the external validation (values reported are averages of the 10 prediction sets).

Table 4: Predefined Training and Prediction Set PLS Results[a]

model	LV	*X* variance (cumulative)	R^2	Q^2	SDEC	SDEP	R^{2ext}	$SDEP^{ext}$
Lig1C	1	94.84	0.29	0.25	0.74	0.76	0.27	0.77
	2	94.08	0.65	0.56	0.51	0.58	0.73	0.54
	3	94.74	0.75	0.58	0.44	0.57	0.64	0.57
Lig3C	1	33.54	0.37	0.27	0.69	0.75	0.50	0.67
	2	55.13	0.46	0.37	0.64	0.70	0.67	0.57
	3	80.05	0.50	0.39	0.62	0.69	0.64	0.59
	4	83.07	0.62	0.45	0.54	0.66	0.62	0.60
	5	87.57	0.68	0.49	0.49	0.62	0.66	0.58
	6	94.53	0.71	0.54	0.47	0.60	0.61	0.59
	7	96.29	0.75	0.55	0.43	0.59	0.63	0.56

[a] LV represents the number of latent variables in the model. The optimal number of LVs is shown in bold. R^2 and Q^2 represent the correlation coefficient of training and LOO. SDEC and SDEP are the standard deviations of calculation and prediction for internal validation. R^{2ext} and $SDEP^{ext}$ are similarly defined applied to the external validation.

Table 5: Comparison between Various 3D-QSAR Methods and Models Generated by SE-COMBINE[a]

method	descriptors	Q^2	LV	compds	predictive R^2	SDEP
CoMFA	2184	0.63	5	72	0.65	0.52
CoMSIA	2184	0.75	9	72	0.84	0.35
MQS matrices	20/72	0.63	8	72	0.75	0.47
fragment QS-SM	15−25/95	0.69	8	69	0.92	0.51
SE-COMBINE Lig1C	477	0.58	3	72	0.64	0.57
SE-COMBINE Lig3C	1389	0.55	7	72	0.63	0.56

[a] The total number of descriptors in each model is given. The LOO cross-validated Q^2, number of latent variables (LV), compounds (compds) in the training set, predictive R^2 for the 16 compound prediction set, and the SDEP.

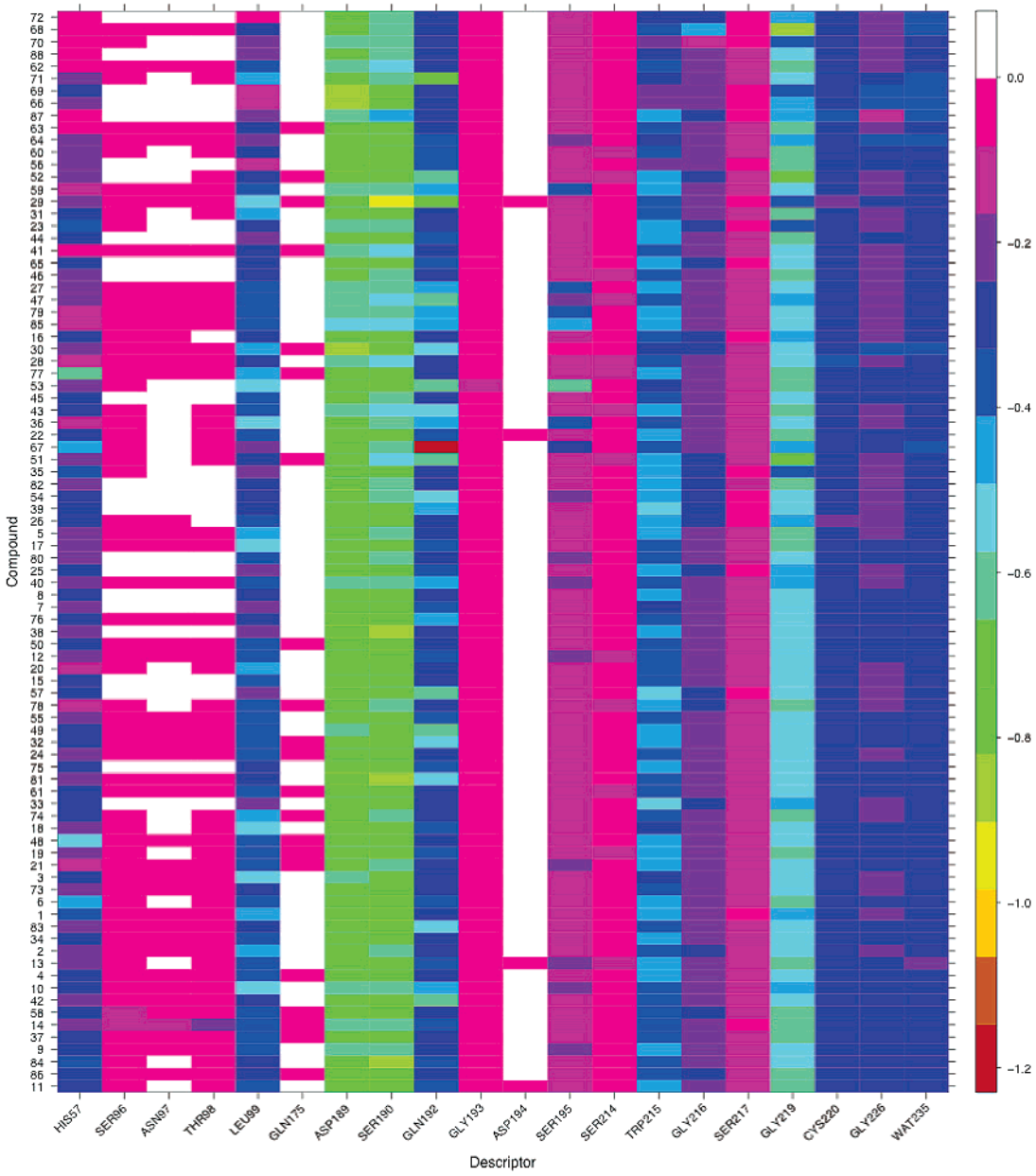


Figure 13. Model Lig1C intermolecular interaction map (IMM) of the important E_{AB} descriptors. The key residues of trypsin that interact with the single fragment ligand (**TAP**) label the *x*-axis. The compounds on the *y*-axis are ordered with respect to activity. The activity decreases from top to bottom. The legend indicates the magnitude of the unscaled descriptor in eV.

The loadings plot of PC 1 and PC 2 is shown in Figure 9.C. There are two clear clusters in the loading plot, the first contains the E_{AB}^{core} and E'_{AB} descriptors and the second all the E_{AB} terms. Importantly, the derivation of the PWD ascribes to E_{AB} the binding information between the fragments. Figure 9.D takes a closer look at the 29 E_{AB} terms. The scores and loadings in a PCA are related: the scores provide the coordinates of the data in the so-called hyper-planes, and the loadings present the direction of each dimension. The loading plot sheds light on the reason clustering in the score plot occurs, because the link between the two plots can be made by comparing the position of the original variables in the loading plot and the position of the compounds in the score plot. His57 is a part of the catalytic triad of trypsin and is prominent in the upper right quadrant of the loading plot. This is in agreement with the score plot where groups with smaller R^2 group were found in that region. The interaction of His57 and the inhibitors would be expected to be greater with a larger R group. Trp215 in the lower right quadrant dominates which is a key residue in the **S3** pocket. Inhibitors with smaller R^1 groups such as **66** and **69** are shown in the lower right quadrant. The residues that lie between the upper quadrants make up the base of the **S1** pocket, which interact with the benzamidine moiety of the inhibitors. There is no obvious relationship between their orientations in the loading plot and the score plot.

Principal Component Analysis: Model Lig3C. This model contains 1389 energy descriptors, and a PCA model was generated where 7 components account for 97% of the variation in the *X* matrix as shown in Figure 10.A. The score plot of PC 1 and 2 is shown in Figure 10.B with two components explaining 71% of the variance in the *X* matrix.

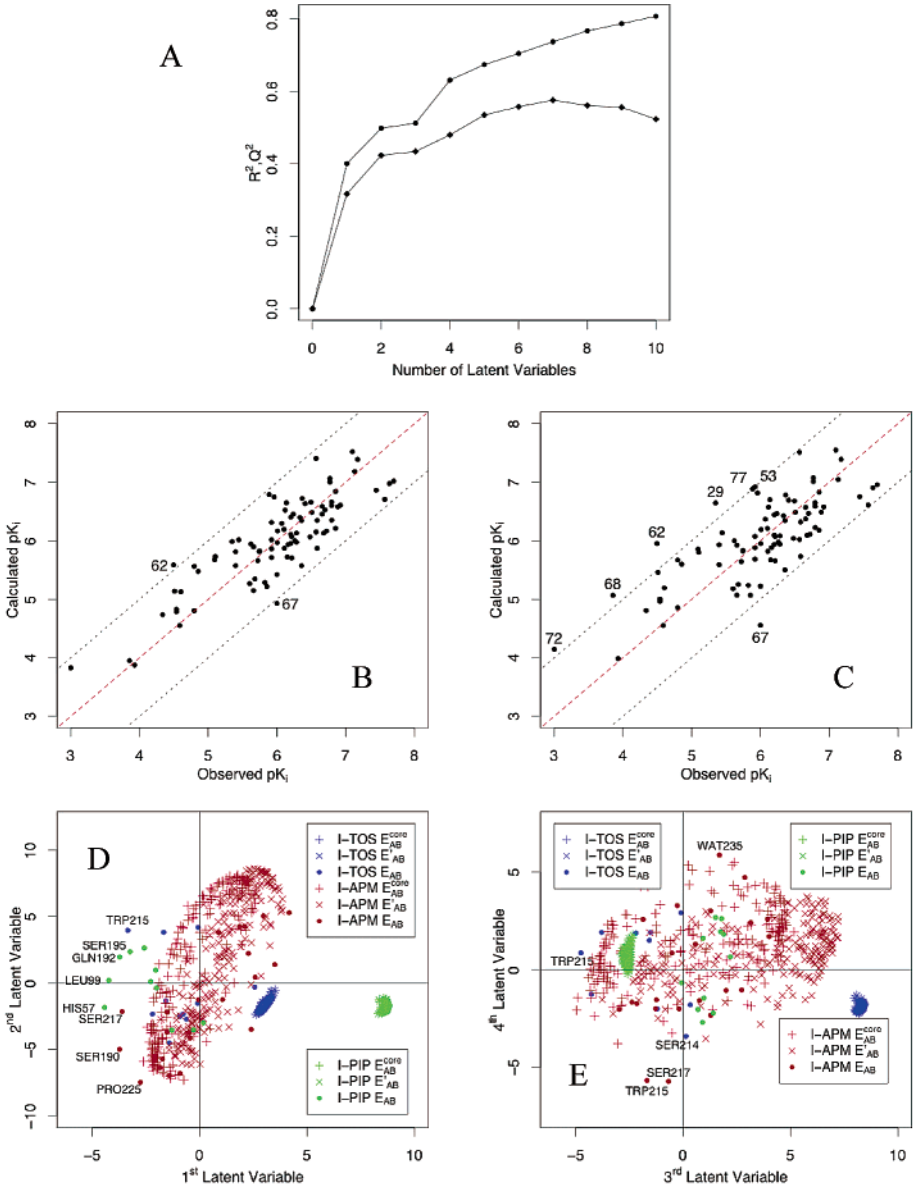


Figure 14. Model Lig3C PLS results. (A) R^2 and Q^2 versus number of latent variables. R^2 is represented as solid circles and Q^2 as solid diamonds. (B) Observed versus calculated pK_i values from internal validation. The red line represents the optimal correlation, while the blue lines are a pK_i unit from the optimal line. (C) Observed versus calculated pK_i values from LOO cross-validation. (D) Loading plot of latent variable 1 versus 2. (E) Loading plot of latent variable 3 versus 4.

Similar to model Lig1C the complexes are differentiated with the R group size playing a large role. Very large R^1 substituents lie to the left of the plot and decrease in size from left to right. The R^2 group sizes decrease from top to bottom. In other words, PC 1 explains the variation in R^1 and PC 2 explains the dissimilarity in R^2. The corresponding loading plot is shown in Figure 10.C, and again similar clustering occurs as model Lig1C. However, a deeper understanding of the key interactions can be constructed, due to the fragmentation of the ligands. After pruning model Lig3B, 11 I-**TOS**, 11 I-**PIP**, and 23 I-**APM** E_{AB} descriptors remained where I denotes any amino acid residue of the protein. The interactions between Trp215, Gln192, and Gly216 and the fragment **TOS** or R^1 are shown in Figure 10.D as dominant contacts. The larger the group at the R^1 position results in a greater interaction with Trp215. His57-, Gln192-, Leu99-, and Ser195-PIP dominates PC 2. Ser195 is a part of the catalytic triad where Leu99 lies between the **S2** and **S3** pockets. Both **S2** and **S3** are large pockets, and the results confirm the optimization of the size and shape of R^1 and R^2. Both interactions between the R groups and the residues of each pocket are dominated by hydrophobic contacts. As a way to demonstrate this, the change in activity by the substitution of 4-methylpiperidide (**4**) by *N*-methyl (**63**) at the R^2 position results in a 2.25 log unit loss in activity.

Partial Least Squares: Model Lig1C. A PLS model is generated to maximally explain the variance in *X* that correlates with *Y*. The statistical quantities of the Lig1C PLS model are shown in Table 3 such as R^2, Q^2, SDEC, and SDEP and the externally validated SDEC and SDEP.

R^2 and Q^2 plots against the number of latent variables (LVs) are shown in Figure 11.A. The R^2 values gradually increase with every additional latent variable as expected; however, the Q^2 value reaches a peak at 3 and tails off.

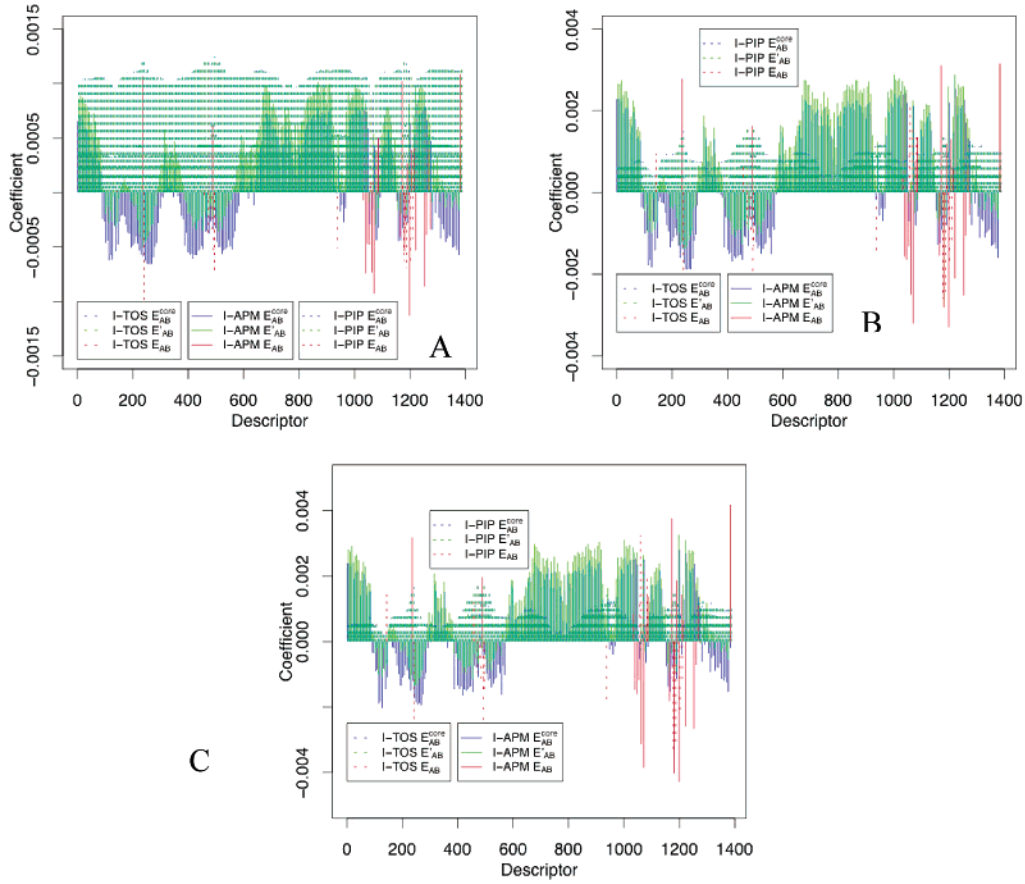


Figure 15. Model Lig3C PLS coefficient plots. (A) Latent variable 1. (B) Latent variable 2. (C) Latent variable 3.

Therefore, the optimal dimensionality of the PLS model involved 3 latent variables. Values higher than 0.5 for R^2 were considered statistically significant. Values greater than 0.4 for Q^2 were viewed as significant, and the optimal relationship between R^2 and Q^2 is the case where $R^2/Q^2 = 1$.[43] Although model Lig1C does not satisfy this equality, the 3 LV model explains 95% of the X matrix and 73% of the Y vector with a Q^2 of 0.59 and an SDEC of 0.45. The externally validated SDEP value of 0.42 is similar to that of the internal validation value, 0.56, and suggests a robust model. The predicted pK_i values are plotted against the experimental values in Figures 11.B, and the values predicted in the LOO cross-validation are shown in Figure 11.C.

When all the complexes were used to build the model, ligands **62** and **77** were presented as outliers (residual pK_i greater than 1 pK_i unit). Compounds **48** and **77** are similar structures where a proton is replaced with a methyl group. Model Lig1C predicts **48** to have a pK_i of 6.511 (6.658) with a residual of 0.147. However, the predicted value of **77** is 6.971 (5.886), a residual of 1.085. The methyl group in **77** is pointing directly into solution, while **48** has the ability to form hydrogen bonds with surrounding water molecules. The inclusion of solvation effects in our modeling would likely improve our ability to accurately model **77**. Compound **62** is also overestimated with a residual of 0.995. Nonetheless, it is considered more important that the predicted pK_i of similar structures **5** (6.328) and **17** (6.544) follows the same trend as that of the experimental values.

Using the LOO cross-validated approach the predicted values **29**, **67**, and **71** can also be identified as outliers. Compounds **29** and **71** are closely related to the high affinity ligand **10**. The order of the predicted pK_i's is incorrect when compared to the experimental values. Compounds **10**, **29**, and **71** are predicted to have activities of 6.514, 6.955, and 5.873, respectively. Compounds **10** and **29** differ in the orientation of the carboxylate group where Ser195 forms a hydrogen bond with **10**, while Gln192 interacts with **29**. The overestimation of the pK_i for **29** is probably due to the strong interaction with Gln192 that was "unseen" in the training phase. The distribution of the descriptors in the PLS model is shown in Figure 11.D where LV 1 is predominantly defined by the E^{core}_{AB} and E'_{AB} descriptors and the Trp215 E_{AB} term. Asp194, Pro225, and Gly196 define the second LV. His57 and Leu99 make strong contributions to both LVs.

These dominant interactions can be verified by looking at the PLS coefficient plots in Figure 12 and the Intermolecular Interaction Map (IMM) in Figure 13. The IMM plot highlights the dominant E_{AB} terms between the ligands and the receptor. IMMs allow the medicinal chemist to graphically represent the change in ligand fragment substitutions with an associated change in intermolecular energy at the residue level. For example, the interaction of **77** and **48** with His57 can easily be distinguished. Also the interaction of anthraquinone-2-sulfonyl (ACS) of **14** interacting with Thr98 is highlighted. The optimization of the ACS-like fragment with Thr98 may lead to stronger binding inhibitors. The score plots of the first and second latent variable are shown in Figure 11 (E and F). The compounds in the upper right quadrant are of high affinity, while those in the lower left

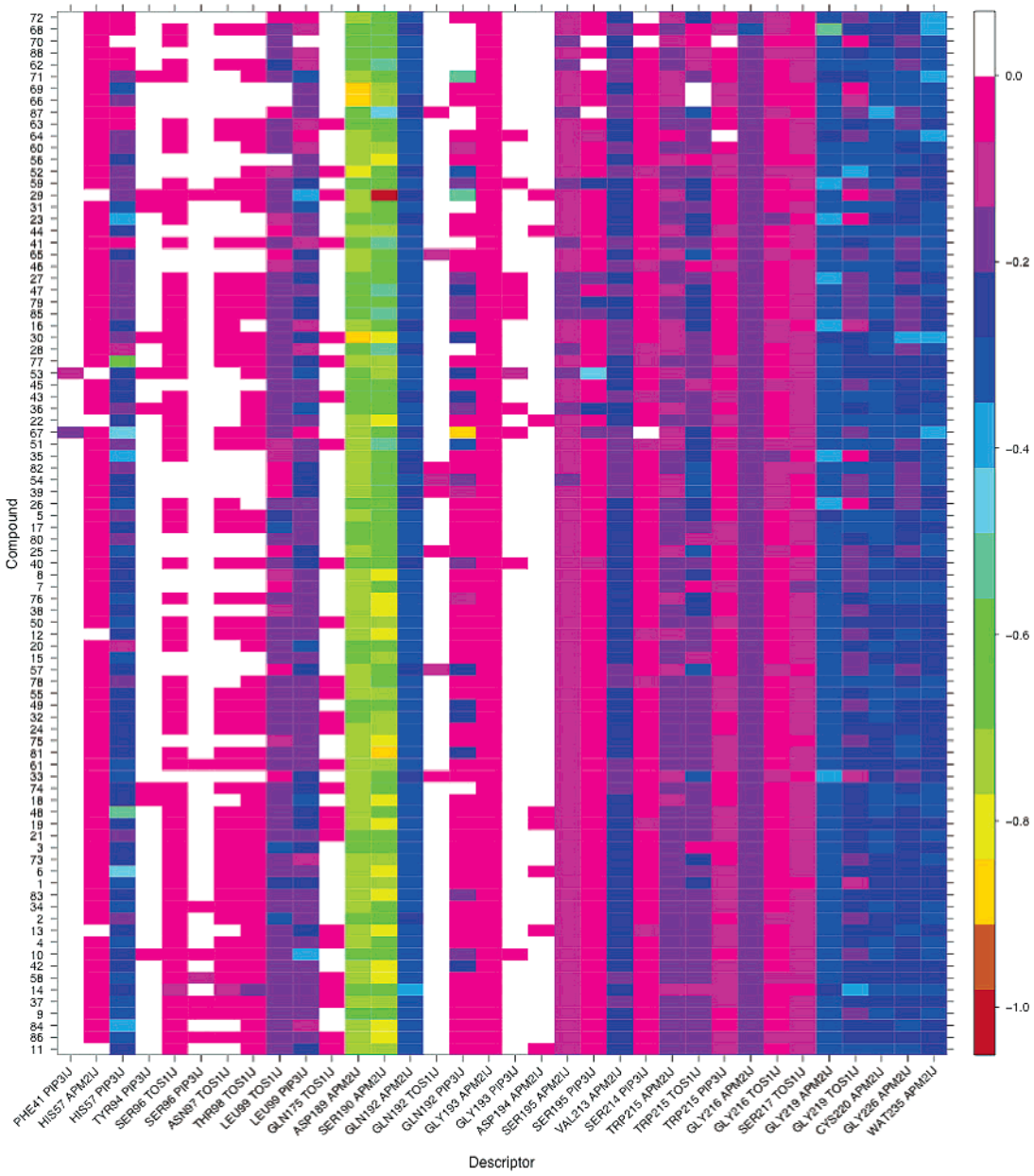


Figure 16. Model Lig3C intermolecular interaction map (IMM) of the important E_{AB} descriptors. The key residues of trypsin that interact with the triple fragment ligand (**APM**, **TOS**, and **PIP**) label the *x*-axis. The compounds on the *y*-axis are ordered with respect to activity. The activity decreases from top to bottom. The legend indicates the magnitude of the unscaled descriptor in eV.

quadrant have the lowest affinity.[44] The compounds incorrectly explained by a latent variable lie on the off-diagonal. Overestimated compounds populate the lower right quadrant, and underestimated compounds occupy the upper left quadrant. The first two LVs account for 94.18% of the variance in descriptor space. There is a greater spread around the optimal line in the score plot of the first LV than the second. The tight binding compounds **11** and **9** are correctly placed in the upper right quadrant of the score plot for LV 2, while the weaker binders such as **72** are also properly placed in the lower left quadrant. This suggests the importance of the E_{AB} descriptors even though they are in the minority. In the case of compound **11**, the electronegative oxygen atoms of the methyl ester on the piperazine group are pointed into **S2**, which is a favorable interaction. The 4-phenyl piperazide group of compound **84** fills the **S2** pocket and is predicted to have a high affinity. Compounds **10** and **71** differ by the stereochemistry of their carboxylate groups. **10** is a high affinity inhibitor, while **71** is a poor inhibitor. **71** is overestimated in the model (see Figure 11.C,F), and the analysis of the score plot confirms this.

Partial Least Squares: Model Lig3C. The statistical results for model Lig3C are shown in Table 3. The development of R^2 and Q^2 can be seen in Figure 14.A. The 7 LV model has an R^2 of 0.74, a Q^2 of 0.58, an SDEC of 0.45, an SDEP of 0.57. This model explains 96% of the *X* matrix and 74% of the activity variation. The external validated SDEP of 0.51 is similar to that of the internal LOO validated value thus suggesting a robust and predictive model. Statistically Lig1C and Lig3C are very similar; however, in terms of interpretability Lig3C is far superior. Lig3C allows the modeler to assess the interaction energy change upon

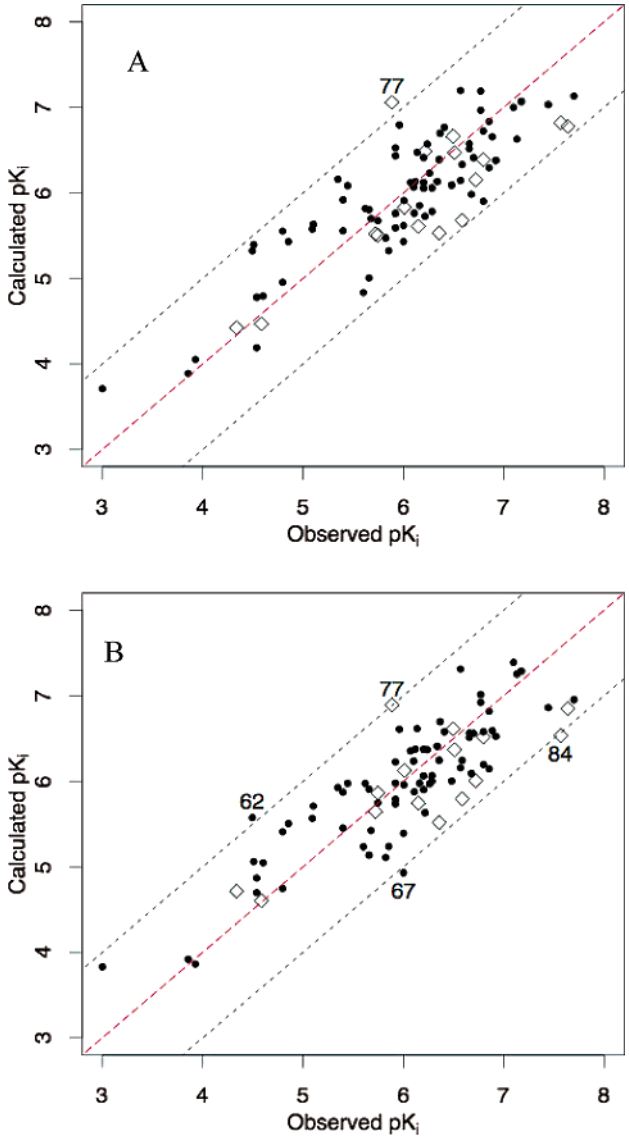


Figure 17. Models Lig1C and Lig3C derived using the training (circles) and prediction (diamonds) sets outlined by Böhm et al. (A) Model Lig1C. (B) Model Lig3C. The red line represents the optimal correlation, while the blue lines are a pK_i unit from the optimal line.

fragment substitution. The predicted pK_i values are plotted against the experimental values in Figure 14.B, and the values predicted by LOO cross-validation are shown in Figure 14.C. The compounds **62** and **67** are outliers in the model. Compounds **29**, **53**, **68**, **72**, and **84** are also outliers in the LOO scheme. The loading plot for LV 1 versus LV 2 and LV3 versus LV 4 are shown in Figure 14D,E. The interactions of Ser195, Gln192, Leu99, and His57 with the fragment **PIP** dominates the first LV, while the interactions of Ser190 and Pro225 with the fragment **APM** dictates the second LV. LV 3 is characterized by the Trp215 interaction with fragment **TOS**, and LV 4 is distinguished by the interaction of Wat235, Ser217, and Trp215 with **APM**. The coefficient plots of the first three latent variables are shown in Figure 15, which complement the loading plots. Similar to the Lig1C model an IMM can help to decipher the reasons for outliers and low affinity of certain inhibitors. The IMM of the important E_{AB} descriptors are shown in Figure 16. The variance in activity of the inhibitors can be predominantly explained by the interactions with the amino acid residues listed above. Interestingly, Asp189 is not significant in the model; however, as shown in the IMM, the interactions with Asp189 are strong and are essential for binding.

Compound **62** presents as an outlier in both SE-COMBINE models. Ligand **62** is overestimated in this model with a residual value of 1.092 pK_i units. Similar to model Lig1C the trend of predicted activity values of **5**, **17**, and **62** are in the same order as experiment. The predicted value for **77** is overestimated compared to **48**, which is a similar result to model Lig1C in both training and cross-validation. The IMM highlights the difference between the two by considering the interaction between His57 and **PIP**. The interaction is stronger for **77** compared to **48** and results in overbinding. The interactions of Gln192 and **PIP** of compounds **71**, **29**, and **67** are stronger compared to the rest of the compounds as shown in the IMM. The activity of **29** is overestimated compared to compounds **10** and **71** due to the strong interaction with Gln192. **53** and **67** interact with Phe41 deep in the **S2** pocket; however, this does not translate into increased activity. The activity values are concentrated between 5.5 and 7 pK_i units, and so the modeling of compound **72** is challenging due to its low activity. **72** is a leverage point, meaning that it has a high influence on the model. Experimentally the addition of a carboxylate group to **78** creating **60** causes a pK_i drop of 1.561. The IMM suggests that the carboxylate interacts favorably with Gln192; however, poorer binding activity results.

Model Lig1C and Lig3C Derived using Predefined Training and Prediction Sets. To test the ability of SE-COMBINE to be used in a real lead optimization situation QSAR models were built using the training and prediction sets outlined by Böhm et al. where the first 72 compounds were placed in the training set and the remaining 16 compounds made up the prediction set. Two different studies have been carried out using this division of the data set. Böhm et al.[35] reported a CoMFA and CoMSIA study in 1999 and Robert et al.[37] described a quantum similarity study in 2000. Again, the Lig1C and Lig3C descriptors were used to construct two QSAR models, and the results of these models are shown in Table 4. The plots of observed versus predicted activity values of both the training and prediction sets for both models are shown in Figure 17. The predicted pK_i residual of compound **77** in model Lig1C is greater than one pK_i unit, while **62**, **67**, and **84** have similar residuals in model Lig3C. A comparison was then made between the different methods used to predict the binding affinity of the 88 compounds to trypsin. On initial inspection the SE-COMBINE methods seems to underperform its nonreceptor based counterparts as shown in Table 5. The CoMSIA approach has the highest Q^2 value of 0.75 and the lowest SDEP. The CoMSIA approach calculates the steric occupancy, partial atomic charges, local hydrophobicity, and hydrogen-bond donor and acceptor properties at each grid point. Currently, SE-COMBINE only considers the electrostatic interaction between the receptor and ligand, and such effects as receptor and ligand desolvation and dispersion are neglected. On the basis of this, it is not an unexpected result

that SE-COMBINE does not perform as well as these methods. The Fragment QS-SM method removed three compounds from the training set; however, the Q^2 is still only 0.69, and it has an SDEP of 0.51 which is similar to the SE-COMBINE models. The QSM methods have a high number of LVs compared to the number of descriptors, and overfitting could be an issue as well.

Conclusion

This research describes the derivation and implementation of a new receptor-based QSAR called SE-COMBINE. The validation of SE-COMBINE was used as an investigation of the interactions between trypsin and a series of inhibitors. The interactions between key residues of the protein and fragments of the ligand were elucidated, and their changes were compared to experimental data. The research shows that SE-COMBINE can be used in a lead optimization scheme in structure-based drug-design. This method allows the chemist to investigate the gain or loss of interaction energy upon fragment substitution. SE-COMBINE was not directly compared to another receptor-based method as a control experiment; however, considering that SE-COMBINE includes effects such as charge transfer and polarization, it is possible that it would outperform its molecular mechanics counterparts, such as COMBINE and MM/PBSA. The models generated using SE-COMBINE were competitive when compared to nonreceptor based QSARs, considering that it uses an incomplete energy function.

The decomposed interaction energies have shed light onto the key interactions between trypsin and a series of benzamidine-based inhibitors. Using current statistical and graphical tools PWD and SE-COMBINE has the potential to be a powerful technique in structure-based drug design. From the detailed analysis of the protein−ligand interactions, predictions of new and improved inhibitors can be made. Model Lig3C highlighted the important ligand fragment−protein residue interactions and thus allows a computational chemist to create hypothetical virtual compounds and predict their activity using the model. For example, the ACS group of **14** could be optimized to enhance the interactions with Thr98 or modify other scaffolds to include this interaction.

SE-COMBINE is not limited to drug design. Problems such as protein-decoy-discrimination, protein stability, and protein metal ion selectivity and in silico protein mutagenesis studies can be targeted using this approach. Investigating such problems with SE-COMBINE can only lead to a better understanding of molecular stability, recognition, selectivity, and ultimately a more complete understanding of molecular interactions.

SE-COMBINE in its current form does not decompose a complete energy function. Recent developments have shown that the solvation free energy of binding can be partitioned using either a Generalized-Born or Poisson−Boltzmann approach. Dispersion effects are neglected in QM methods. These could easily be included in the form of a Lennard-Jones ($(1/R^6)$) potential. The attractive part of the LJ potential lends itself to be partitioned; in fact it is widely used in MM potentials. The entropy term of the master binding equation thus remains. Entropy is not an intermolecular interaction, and so it is almost impossible to have a complete pairwise function. However, it could easily be added to the potential function without being partitioned. Hence, future work will add solvation, dispersion, and entropy components to SE-COMBINE which, in effect, would result in a pairwise QMSCORE.[25]

Acknowledgment. The authors are grateful to Gerhard Klebe for providing the trypsin data set and to Rajarshi Guha for many useful discussions. We thank the NSF (MCB-0211639) and the NIH (GM 44974) for financial support of this research. The helpful comments by the anonymous reviewers are also acknowledged.

References

(1) Ferrara, P.; Gohlke, H.; Price, D. J.; Klebe, G.; Brooks, C. L. Assessing scoring functions for protein−ligand interactions. *J. Med. Chem.* **2004**, *47* (12), 3032−3047.

(2) Hansch, C. A Quantitative Approach to Biochemical Structure−Activity Relationships. *Acc. Chem. Res.* **1969**, *2*, 232−239.

(3) Cramer, R. D., III; Patterson, D. E.; Bunce, J. D. Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959−5967.

(4) Wade, R. C.; Ortiz, A. R.; Gago, F., Comparative binding energy analysis. *Perspect. Drug Discovery Des.* **1998**, *9*−*11*, 19−34.

(5) Ortiz, a. R.; Pisabarro, M. T.; Gago, F.; Wade, R. C. Prediction of Drug-Binding Affinities by Comparative Binding-Energy Analysis. *J. Med. Chem.* **1995**, *38* (14), 2681−2691.

(6) Kuhn, B.; Kollman, P. A. Binding of a Diverse Set of Ligands to Avidin and Streptavidin: An Accurate Quantitative Prediction of their Relative Affinites by Combination of Molecular Mechanics and Continuum Solvation Models. *J. Med. Chem.* **2000**, *43* (20), 3786−3791.

(7) Wang, W.; Lim, W. A.; Jakalian, A.; Wang, J.; Wang, J.; Luo, R.; Bayly, C. I.; Kollman, P. A., An Analysis of the Interactions between the Sem-5 SH3 Domain and its Ligands Using Molecular Dynamics, Free Energy Calculations, and Sequence Analysis. *J. Am. Chem. Soc.* **2001**, *123*, 3986−3994.

(8) Wold, S.; Trygg, J.; Berglund, A.; Antti, H. Some recent developments in PLS modeling. *Chemom. Intell. Lab. Syst.* **2001**, *58* (2), 131−150.

(9) Wold, S.; Sjostrom, M.; Eriksson, L. PLS-regression: a basic tool of chemometrics. *Chemom. Intell. Lab. Syst.* **2001**, *58* (2), 109−130.

(10) Pople, J. A.; Beveridge, D. L. In *Approximate Molecular Orbital Theory*; McGraw-Hill: New York, 1970.

(11) van der Vaart, A.; Gogonea, V.; Dixon, S. L.; Merz, K. M., Jr. Linear Scaling Molecular Orbital Calculations of Biological Systems Using the Semiempirical Divide and Conquer Method. *J. Comput. Chem.* **2000**, *21*, 1494−1504.

(12) Dixon, S. L.; Merz, K. M., Jr. Fast, Accurate Semiempirical Molecular Orbital Calculations for Macromolecules. *J. Chem. Phys.* **1997**, *107*, 879−893.

(13) Dixon, S. L.; Merz, K. M., Jr. Semiempirical Molecular Orbital Calculations with Linear System Size Scaling. *J. Chem. Phys.* **1996**, *104*, 6643−6649.

(14) Yang, W.; Lee, T.-S. A Density-matrix Divide-and-conquer Approach for Electronic Structure Calculations of Large Molecules. *J. Chem. Phys.* **1995**, *103* (13), 5674−5678.

(15) Li, X. P.; Nunes, R. W.; Vanderbilt, D. Density-Matrix Electronic-Structure Method with Linear System-Size Scaling. *Phys. Rev. B* **1993**, *47* (16), 10891−10894.

(16) Stewart, J. J. P. Application of localized molecular orbitals to the solution of semiempirical self-consistent field equations. *Int. J. Quantum Chem.* **1996**, *58* (2), 133−146.

(17) Dixon, S. L.; van der Vaart, A.; Gogonea, V.; Vincent, J. J.; Brothers, E. N.; Suárez, D.; Westerhoff, L. M.; Merz, K. M., Jr. *DIVCON99*, The Pennsylvania State University: 1999.

(18) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. AM1: A New General Purpose Quantum Mechanical Molecular Model. *J. Am. Chem. Soc.* **1985**, *107*, 3902−3909.

(19) Stewart, J. J. P. Optimization of Parameters for Semiempirical Methods I. Method. *J. Comput. Chem.* **1989**, *10*, 209−220.

(20) Stewart, J. J. P. Optimization of Parameters for Semiempirical Methods II. Applications. *J. Comput. Chem.* **1989**, *10*, 221−264.

(21) Thiel, W.; Voityuk, A. A. Extension of MNDO to d Orbitals: Parameters and Results for the Second-Row Elements and for the Zinc Group. *J. Phys. Chem.* **1996**, *100*, 616−626.

(22) Dewar, M. J. S.; Thiel, W. Ground States of Molecules. 38. The MNDO method. Approximations and Parameters. *J. Am. Chem. Soc.* **1977**, *99* (15), 4899−4907.

(23) Repasky, M. P.; Chandrasekhar, J.; Jorgensen, W. L. PDDG/PM3 and PDDG/MNDO: Improved semiempirical methods. *J. Comput. Chem.* **2002**, *23* (16), 1601−1622.

(24) Raha, K.; Merz, K. M., Jr. A Quantum Mechanics Based Scoring Function: Study of Zinc-ion Mediated Ligand Binding. *J. Am. Chem. Soc.* **2004**, *126*, 1020−1021.

(25) Raha, K.; Merz, K. M., Jr. Large-scale validation of a quantum mechanics based scoring function: predicting the binding affinity and the binding mode of a diverse set of protein−ligand complexes. *J. Med. Chem.* **2005**, *48* (14), 4558−75.

(26) Raha, K.; van der Vaart, A. J.; Riley, K. E.; Peters, M. B.; Westerhoff, L. M.; Kim, H.; Merz, K. M., Jr. Pairwise Decomposition of Residue Interaction Energies Using Semiempirical Quantum Mechanical Methods in Studies of Protein−Ligand Interaction. *J. Am. Chem. Soc.* **2005**, *127* (18), 6583−6594.

(27) Wang, B.; Raha, K.; Merz, K. M., Jr. Pose Scoring by NMR. *J. Am. Chem. Soc.* **2004**, *126* (37), 11430−11431.

(28) Davie, E. W.; Fujikawa, K.; Kisiel, W. The Coagulation Cascade - Initiation, Maintenance, and Regulation. *Biochemistry* **1991**, *30* (43), 10363−10370.

(29) Silverman, R. B. *The organic chemistry of enzyme-catalyzed reactions*; Academic Press: San Diego, 2002.

(30) Turk, D.; Sturzebecher, J.; Bode, W. Geometry of Binding of the N-Alpha-Tosylated Piperidides of Meta-Amidino-Phenylalanine, Para-Amidino-Phenylalanine and Para-Guanidino-Phenylalanine to Thrombin and Trypsin - X-ray Crystal-Structures of Their Trypsin Complexes and Modeling of Their Thrombin Complexes. *FEBS Lett.* **1991**, *287* (1−2), 133−138.

(31) Dullweber, F.; Stubbs, M. T.; Musil, D.; Sturzebecher, J.; Klebe, G. Factorising ligand affinity: A combined thermodynamic and crystallographic study of trypsin and thrombin inhibition. *J. Mol. Biol.* **2001**, *313* (3), 593−614.

(32) Mares-Guia, M.; Shaw, E. Studies on the active center of trypsin; the binding of amidines and guanidines as models of the substrate side chain. *J. Biol. Chem.* **1965**, *240*, 1579−1585.

(33) Sturzebecher, J.; Prasa, D.; Hauptmann, J.; Vieweg, H.; Wilkstrom, P. Synthesis and structure−activity relationships of potent thrombin inhibitors: Piperazides of 3-amidinophenylalanine. *J. Med. Chem.* **1997**, *40* (19), 3091−3099.

(34) Sturzebecher, J.; Prasa, D.; Wikstrom, P.; Vieweg, H. Structure−Activity-Relationships of Inhibitors Derived from 3-Amidinophenylalanine. *J. Enzymol. Inhib.* **1995**, *9* (1), 87−99.

(35) Böhm, M.; Sturzebecher, J.; Klebe, G. Three-dimensional quantitative structure−activity relationship analyses using comparative molecular field analysis and comparative molecular similarity indices analysis to elucidate selectivity differences of inhibitors binding to trypsin, thrombin, and factor Xa. *J. Med. Chem.* **1999**, *42* (3), 458−477.

(36) Klebe, G. Comparative Molecular Similarity Indices: CoMSIA. In *3D QSAR in Drug Design*; Kubinyi, H., Folkers, G., Martin, Y. C., Eds.; Kluwer Academic Publishers: Great Britain, 1998; Vol. 3, p 87.

(37) Robert, D.; Amat, L.; Carbo-Dorca, R. Quantum similarity QSAR: Study of inhibitors binding to thrombin, trypsin and factor Xa, including a comparison with CoMFA and CoMSIA methods. *Int. J. Quantum Chem.* **2000**, *80* (3), 265−282.

(38) Murcia, M.; Ortiz, A. R. Virtual screening with flexible docking and COMBINE-based models. Application to a series of factor Xa inhibitors. *J. Med. Chem.* **2004**, *47* (4), 805−820.

(39) Cornell, W. D.; Cieplak, P.; Baylay, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. A Second Generation Force Field For the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.* **1995**, *117*, 5179−5197.

(40) R *R: A Language and Environment for Statistical Computing*, *2.0.1*; R Development Core Team: R Foundation for Statistical Computing: Vienna, Austria, 2005.

(41) Mevik, B. H.; Cederkvist, H. R. Mean squared error of prediction (MSEP) estimates for principal component regression (PCR) and partial least squares regression (PLSR). *J. Chemom.* **2004**, *18* (9), 422−429.

(42) Cundari, T. R.; Sarbu, C.; Pop, H. F. Robust fuzzy principal component analysis (FPCA). A comparative study concerning interaction of carbon−hydrogen bonds with molybdenum-oxo bonds. *J. Chem. Inf. Comput. Sci.* **2002**, *42* (6), 1363−1369.

(43) Golbraikh, A.; Tropsha, a., Beware of q(2)! *J. Mol. Graphics* **2002**, *20*, (4), 269−276.

(44) Stanton, D. T. On the physical interpretation of QSAR models. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (5), 1423−1433.

CT050284J

JCTC Journal of Chemical Theory and Computation

Density-Fitting Method in Symmetry-Adapted Perturbation Theory Based on Kohn−Sham Description of Monomers

Rafał Podeszwa,* Robert Bukowski, and Krzysztof Szalewicz

Department of Physics and Astronomy, University of Delaware, Newark, Delaware 19716

Received December 2, 2005

Abstract: We present a new implementation of symmetry-adapted perturbation theory of intermolecular interactions based on Kohn−Sham description of monomers. With density-fitting of molecular integrals, the scaling of the computational cost of the method is reduced from the sixth to the fifth power of the system size. Computational requirements of some operations scaling as the fifth power have also been significantly reduced. The new method allows an accurate treatment of molecules consisting of as many as a few dozen of atoms, using both nonhybrid and hybrid density functionals.

I. Introduction

Structure and properties of various atomic and molecular systems, from clusters to condensed phases to biological molecules, are determined by weak intermolecular interactions, also referred to as van der Waals interactions. These effects have been successfully studied ab initio, both within the supermolecular framework, using the coupled-cluster method and many-body perturbation theory, and perturbatively in intermolecular interaction operator V, using methods such as symmetry-adapted perturbation theory (SAPT).[1−3] Unfortunately, the computational cost of the wave function-based approaches increases prohibitively fast with system size N, as $O(N^7)$ for methods including triple excitations. Density functional theory (DFT) is much less time-consuming; however, the existing versions of DFT, when applied within the supermolecular approach, fail to reproduce the dispersion interaction, an important part of the van der Waals force. This problem is due to the fact that dispersion forces result from long-range correlations between electrons, whereas the current exchange-correlation potentials model only local correlation effects.

Another approach to the calculations of interaction energies is based on SAPT but utilizes the description of the interacting monomers in terms of Kohn−Sham (KS) orbitals and orbital energies, a method now called SAPT(KS). The predictions of the original version of this approach, proposed by Williams and Chabalowski,[4] were rather poor, and these authors conjectured that the reason could be the wrong asymptotic behavior of the KS electron densities. In a subsequent development, Misquitta and one of the present authors[5] and independently Hesselmann and Jansen[6,7] have shown that indeed if the asymptotic behavior is corrected, the accuracy of the electrostatic, exchange, and induction terms greatly improves. Only the dispersion component was still inaccurate. This problem was found[8,9] to be due to the use of a formula asymptotically related to uncoupled dynamic polarizabilities. When, instead, the dispersion energies were calculated from frequency-dependent density susceptibility (FDDS) functions, also referred to as propagators, obtained from the time-dependent DFT (TD-DFT) theory at the coupled Kohn−Sham (CKS) level, the results became very accurate. The SAPT approach based on asymptotically corrected KS calculations and on CKS FDDS's was proposed in ref 10 and referred to as SAPT(DFT). The method can be shown to be potentially exact for all major components of the interaction energy (asymptotically for exchange interactions) in the sense that these components would be exact if the DFT description of the monomers were exact.[5,8,11] Applications to a number of small dimers have shown that SAPT(DFT) provides surprisingly accurate individual interaction components, often more accurate than the standard SAPT at the currently programmed level.[10,11] Applications to larger dimers were presented in refs 12 and 13.

10.1021/ct050304h CCC: $33.50
Published on Web 02/18/2006

The nominal scaling of SAPT(KS) is $O(N^5)$ and that of SAPT(DFT) is $O(N^6)$, in both cases significantly better than the $O(N^7)$ scaling of the regular SAPT. Despite this better scaling, it was not feasible to apply SAPT(DFT) to very large systems, e.g., the ones of biological interest, since the $O(N^6)$ scaling is still too steep. Also, some $O(N^5)$ terms of SAPT(DFT) were time-consuming. In ref 8, a partial solution to this problem was proposed based on the density-fitting (also called the resolution of identity) technique,[14−19] which allowed for reducing the scaling of the CKS dispersion energy calculations from $O(N^6)$ to $O(N^3)$. However, the construction of the TD-DFT propagators present in the expression for this energy still required $O(N^6)$ operations. If monomer-centered basis sets were employed, the propagators could be obtained just once for each monomer and then reused (after appropriate translations and rotations) for all dimer geometries. Thus, the construction of the propagators would only be a one-time expense, insignificant compared to the computational effort of obtaining the whole potential energy surface. However, for sufficiently large systems, or when only a single point on the surface is needed, calculations of the propagators could still be the bottleneck. Moreover, the convergence of the dispersion energy is much faster if the propagators are expanded in dimer-centered basis sets rather than in the monomer-centered ones.

Recently, Hesselmann et al.[20] presented an implementation of a method similar to SAPT(DFT), referred to by them as DFT-SAPT. The density-fitting techniques allowed these authors to reduce the scaling of the cost of calculating the CKS propagators to $O(N^4)$ [with $O(N^5)$ overall scaling of the full interaction energy calculation]. However, this approach cannot be applied to hybrid functionals, and the authors of ref 20 suggested using an approximate expression for the Hartree−Fock exchange term,[21] which significantly increased the computational cost. Moreover, the formulation of ref 20 is valid only if all calculations are done in the full basis set of the dimer, which may not be optimal for larger systems.

Recently, we have proposed a new algorithm for calculating the CKS propagators and dispersion energies based on density fitting that can be used with all functionals, including the hybrid ones.[22] The method scales as $O(N^5)$ for hybrid functionals and is equivalent to the formalism of ref 20 for the nonhybrid ones.

In this paper we present a complete implementation of SAPT(DFT) based on density fitting, more general than the implementation of ref 20. Also, in contrast to the latter formulation, all interaction energy components that do not depend on the CKS propagators are evaluated from molecular orbitals, using the standard expressions available in the SAPT2002 code.[23] Apart from the CKS propagators, the density fitting is also used to speed up the transformation that generates molecular integrals needed in these expressions. The outcome is a method with the overall scaling of $O(N^5)$, applicable to both nonhybrid and hybrid functionals, and capable of utilizing both dimer- and monomer-centered basis sets, including the so-called "monomer-centered-plus" bases of ref 24. The precise definition of the SAPT(DFT) approach is presented in section II. Section III discusses the density-fitting approximation used to simplify the transformation of two-electron integrals and CKS calculations. The details of the implementation and an analysis of computational costs are presented in section IV, followed by a discussion in section V of the results obtained for some model systems. Section VI contains conclusions.

II. SAPT(DFT) Method

The SAPT(DFT) method has its roots in the wave function-based SAPT, described in detail in a number of reviews.[1−3] In SAPT, the total Hamiltonian of the dimer *AB* is partitioned as

$$H = F_A + F_B + V + W_A + W_B \tag{1}$$

where F_X and W_X are the Fock operator and the intramonomer correlation operator, respectively, of monomer X ($W_X = H_X - F_X$ with H_X denoting the full Hamiltonian of monomer X), and V is the intermolecular interaction operator. A perturbation theory, starting from the product of Hartree−Fock (HF) determinants of the monomers as the zero-order wave function, gives then the interaction energy in the form of an expansion

$$E_{int} = \sum_{i=1,j=0} (E_{pol}^{(ij)} + E_{exch}^{(ij)}) \tag{2}$$

where the indices i and j denote orders with respect to the operators V and $W = W_A + W_B$, respectively. The polarization terms (with subscript "pol") result from the standard Rayleigh−Schrödinger perturbation theory, whereas the exchange terms (with subscript "exch") arise from antisymmetrization of the dimer wave function in each order. The polarization corrections of the first order in V describe the electrostatic interactions between unperturbed monomers and are therefore denoted by $E_{elst}^{(1j)}$. The second-order corrections can be split into the induction and dispersion components, $E_{pol}^{(2j)} = E_{ind}^{(2j)} + E_{disp}^{(2j)}$ and $E_{exch}^{(2j)} = E_{exch-ind}^{(2j)} + E_{exch-disp}^{(2j)}$. In most applications, it is sufficient to truncate the expansion 2 at the second order in V. If, in addition, all intramonomer correlation corrections are neglected, one obtains the following approximation to the interaction energy (termed SAPT0 without response)

$$E_{int} = E_{elst}^{(10)} + E_{exch}^{(10)} + E_{ind}^{(20)} + E_{disp}^{(20)} + E_{exch-ind}^{(20)} + E_{exch-disp}^{(20)} \tag{3}$$

All terms on the right hand side of eq 3 can be expressed in terms of integrals over HF molecular orbitals of the monomers and orbital energies. In ref 4, Williams and Chabalowski proposed to replace the HF orbitals and energies in these expressions by the ones obtained from DFT Kohn−Sham calculations, hoping that this would compensate for the neglect of intramonomer correlation in the approximation of eq 3. Formally, such an approach corresponds to splitting the dimer Hamiltonian as

$$H = K_A + K_B + V + W_A^{KS} + W_B^{KS} \tag{4}$$

(with K_X denoting the Kohn−Sham operator of monomer X

and $W_X^{KS} = H_X - K_X$) and truncating the resulting perturbation theory at zeroth order in W_X^{KS}.

The original proposal of ref 4 suffered from two fundamental problems. First, most of the commonly used exchange-correlation potentials v_{xc} do not exhibit the proper asymptotic behavior $v_{xc}(\mathbf{r}) \rightarrow -1/r + I + \epsilon_{HOMO}$, where I is the ionization potential and ϵ_{HOMO} is the highest occupied molecular orbital eigenvalue. As it was pointed out in refs 4−7, meaningful interaction energies can only be obtained after an asymptotic correction is applied to the exchange-correlation potential.

The second problem with the formulation of ref 4 was that the expression for $E_{disp}^{(20)}$ evaluated with KS orbitals and orbital energies does not correctly reproduce the dispersion energy. To remedy this flaw, it has been proposed[8,9] to make use of the generalized Casimir-Polder expression,[25−27] relating the dispersion energy to frequency-dependent density susceptibilities, $\alpha_X(\mathbf{r}, \mathbf{r}'|u)$, $X = A, B$

$$E_{disp}^{(2)} = -\frac{1}{2\pi}\int_0^\infty du \int \alpha_A(\mathbf{r}_1,\mathbf{r}'_1|\, iu)\alpha_B(\mathbf{r}_2,\mathbf{r}'_2|\, iu)\frac{d\mathbf{r}_1 d\mathbf{r}_2}{|\mathbf{r}_1 - \mathbf{r}_2|}\frac{d\mathbf{r}'_1 d\mathbf{r}'_2}{|\mathbf{r}'_1 - \mathbf{r}'_2|} \quad (5)$$

where the FDDSs are computed at imaginary frequencies iu. The expression 5 gives the exact second-order dispersion energy as long as the FDDSs are exact. Within the DFT framework, accurate FDDSs can be computed in the CKS approach. Using these FDDSs in eq 5, one obtains a quantity we shall refer to as $E_{disp}^{(2)}$(CKS), which was shown[8−10] to provide a very good approximation to the dispersion energy. Likewise, the induction energy can be computed using the expression[28]

$$E_{ind}^{(2)} = \frac{1}{2}\int \omega_B(\mathbf{r})\alpha_A(\mathbf{r},\mathbf{r}'|0)\omega_B(\mathbf{r}')d\mathbf{r}d\mathbf{r}' + \frac{1}{2}\int \omega_A(\mathbf{r})\alpha_B(\mathbf{r},\mathbf{r}'|\,0)\omega_A(\mathbf{r}')d\mathbf{r}d\mathbf{r}' \quad (6)$$

where ω_X denotes the electrostatic potential generated by monomer X:

$$\omega_X(\mathbf{r}) = \int \frac{\rho_X(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|}d\mathbf{r}' + V_{nuc,X}(\mathbf{r}) \quad (7)$$

In the equation above, $\rho_X(\mathbf{r})$ is the electron density of monomer X, and $V_{nuc,X}(\mathbf{r})$ is the nuclear potential of X. This expression gives the exact second-order induction energy as long as $\rho_X(\mathbf{r})$ and FDDSs at zero frequency are exact. Thus, eqs 5 and 6 together give the exact second-order polarization energy. If expression 6 is computed with uncoupled KS FDDSs and the KS electron densities, the result is equivalent to the calculation of $E_{ind}^{(20)}$ with KS orbitals and orbital energies. If instead the FDDSs at the CKS level are used, one obtains a quantity which we shall refer to as $E_{ind}^{(2)}$(CKS). The KS electron densities are always obtained from asymptotically corrected calculations.

The total SAPT(DFT) interaction energy (up to second order in V) can now be defined as[10]

$$E_{int}^{SAPT(DFT)} = E_{elst}^{(1)}(KS) + E_{exch}^{(1)}(KS) + E_{ind}^{(2)}(CKS) + \tilde{E}_{exch-ind}^{(2)}(CKS) + E_{disp}^{(2)}(CKS) + \tilde{E}_{exch-disp}^{(2)}(CKS) \quad (8)$$

where the terms with CKS label result from the coupled Kohn−Sham approach, whereas the terms labeled KS are obtained by using Kohn−Sham orbitals and orbital energies to compute the corresponding quantities on the right hand side of eq 3. Since the exact expressions for the exchange-induction and exchange-dispersion corrections at the CKS level are not known, we use approximations to these quantities obtained by scaling their KS counterparts:

$$\tilde{E}_{exch-ind}^{(2)}(CKS) = E_{exch-ind}^{(2)}(KS) \times \frac{E_{ind}^{(2)}(CKS)}{E_{ind}^{(2)}(KS)} \quad (9)$$

$$\tilde{E}_{exch-disp}^{(2)}(CKS) = E_{exch-disp}^{(2)}(KS) \times \frac{E_{disp}^{(2)}(CKS)}{E_{disp}^{(2)}(KS)} \quad (10)$$

The approximate nature of these expressions is indicated by the tilde sign. The accuracy of these approximations was tested on small benchmark systems and was shown to be adequate.[10,11]

The SAPT(DFT) interaction energy of eq 8 has been shown in ref 10 to provide a very good approximation to the most accurate available values computed by wave function-based methods. The success of the method can be attributed mostly to the ability of DFT to accurately reproduce molecular densities and response properties. Since the expensive intramonomer correlation terms are avoided [in particular, a SAPT(KS) calculation includes only the expressions given in eq 3], the computational cost of SAPT(DFT) is lower than that of the conventional ab initio methods which have to include these terms to achieve acceptable accuracies. In SAPT(KS), once the integrals over molecular orbitals are available, the computation of all terms in eq 3 takes virtually negligible time compared to the regular SAPT calculation at the complete currently available level of theory. This is due to the $O(N^5)$ vs $O(N^7)$ scaling of the two methods, and in addition to the fact that for $E_{exch-disp}^{(2)}$, the only correction in eq 3 that scales as $O(N^5)$, the precise scaling is $O(o^3v^2)$, with o and v denoting the numbers of occupied and virtual orbitals, and in most cases $o \ll v$. In consequence, SAPT(KS) calculations are dominated by the integral transformation which, for the corrections needed, scales as $O(on^4)$, where $n = o + v$. In SAPT(DFT), however, there are three steps which, for most systems, are more time-consuming than the transformation. (*a*) The construction of the TD-DFT matrices which requires a numerical evaluation of four index integrals involving the derivative of v_{xc} and scales as $O(o^2v^2g)$, where g is the number of integration points. Since g increases with the system size, this integration is an $O(N^5)$ process. (*b*) The evaluation of the TD-DFT propagators, which requires multiplications and inversions of large matrices and scales as $O(N^6)$. (*c*) The calculation of the CKS dispersion energy from the propagators. These three bottlenecks were addressed in refs 8, 10, 20, and 22 and sped up by using density fitting techniques and iterative matrix inversion methods, resulting in scalings of at the most

$O(N^5)$ and greatly reduced prefactors. With these improvements, the integral transformation becomes the most time-consuming step of a SAPT(DFT) calculation for a wide range of systems. In the following sections, we will show how the cost of the transformation can be reduced using density fitting ideas. We will also describe changes that had to be implemented to use the asymptotic correction with the DALTON[29] set of computer codes and develop the density fitting method for the CKS induction energy.

III. Density Fitting Approximation

Two-electron integrals occurring in electronic structure theory can be written in terms of generalized densities, defined as

$$\rho_{ij}(\mathbf{r}) = \phi_i(\mathbf{r})\phi_j(\mathbf{r}) \tag{11}$$

where ϕ_i and ϕ_j are any molecular orbitals of the system considered. The idea of density fitting is to approximate the density ρ_{ij} by

$$\tilde{\rho}_{ij}(\mathbf{r}) = \sum_K^{N_{\text{aux}}} D_K^{ij}\chi_K(\mathbf{r}) \tag{12}$$

where χ_K, $K = 1,\ldots, N_{\text{aux}}$, are auxiliary (fitting) basis functions, usually atom-centered Gaussian orbitals. We assume for simplicity that the auxiliary basis set is identical for all the densities, but, in general, it may depend on i, j. The error introduced by the fitting can be quantified in terms of the functional

$$\Delta_{ij} = \int \mathrm{d}\mathbf{r}_1\mathrm{d}\mathbf{r}_2\,[\rho_{ij}(\mathbf{r}_1) - \tilde{\rho}_{ij}(\mathbf{r}_1)][\rho_{ij}(\mathbf{r}_2) - \tilde{\rho}_{ij}(\mathbf{r}_2)]w(\mathbf{r}_1,\mathbf{r}_2) \tag{13}$$

where $w(\mathbf{r}_1,\mathbf{r}_2)$ is a weight factor. In our implementation, we set $w(\mathbf{r}_1,\mathbf{r}_2) = 1/|\mathbf{r}_1 - \mathbf{r}_2|$, as recommended in ref 16 for fitting electron repulsion integrals. The fit coefficients D_K^{ij} are obtained by minimizing the functional 13, which leads to the following expression

$$D_K^{ij} = \sum_L [\mathbf{J}^{-1}]_{KL}(ij|\,L) \tag{14}$$

where

$$(ij|L) = \int \mathrm{d}\mathbf{r}_1\mathrm{d}\mathbf{r}_2\,\frac{\phi_i(\mathbf{r}_1)\phi_j(\mathbf{r}_1)\chi_L(\mathbf{r}_2)}{|\mathbf{r}_1 - \mathbf{r}_2|} \tag{15}$$

$$J_{LK} = \int \mathrm{d}\mathbf{r}_1\mathrm{d}\mathbf{r}_2\,\frac{\chi_L(\mathbf{r}_1)\chi_K(\mathbf{r}_2)}{|\mathbf{r}_1 - \mathbf{r}_2|} \tag{16}$$

IV. Implementation

A complete SAPT(DFT) calculation for one dimer geometry consists of several steps implemented as separate programs. First, the Kohn−Sham calculations are performed for both monomers, followed by the integral transformation. The transformed integrals are then used to generate the CKS propagators and to obtain the induction and dispersion energies. Finally, the standard code from the SAPT2002 suite[23] is invoked to compute corrections labeled "KS" in eqs 8−10. In the following sections, each of these steps is discussed in more detail.

A. Asymptotically Corrected Kohn−Sham Calculation. In the current version of SAPT(DFT) codes, the Kohn−Sham orbitals and orbital energies for monomers are obtained from the DALTON program.[29] The previous version was interfaced with the CADPAC program[30] which included the asymptotic correction in Kohn−Sham calculations. We have implemented this correction in DALTON in the following way. Using the density $\rho(\mathbf{r})$ from uncorrected Kohn−Sham calculations, the Fermi-Amaldi asymptotic potential[31] is computed as

$$v_{\text{xc,FA}}(\mathbf{r}) = -\frac{1}{N_{\text{el}}}\int \frac{\rho(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|}d\mathbf{r}' \tag{17}$$

where N_{el} denotes the number of electrons. This potential is shifted to obtain the final asymptotic form

$$v_{\text{xc,as}}(\mathbf{r}) = v_{\text{xc,FA}}(\mathbf{r}) + I + \epsilon_{\text{HOMO}} \tag{18}$$

The ionization potential I can be taken from experiment or from a separate ab initio or DFT calculation. The splicing scheme of Tozer and Handy[31] is then used to connect $v_{\text{xc,as}}(\mathbf{r})$ with the standard short-range part. For the splicing scheme, we used Bragg radii factors of 3.0 and 4.0 as recommended in ref 32. The Fermi-Amaldi asymptotic potential will not represent well the true asymptotic potential for large molecules, in particular for long polymers. It has been shown in ref 11 that the asymptotic correction improves the results for the water dimer, but its effects are small for the carbon dioxide dimer. For larger molecules to which SAPT(DFT) has been applied, there are no sufficiently accurate benchmarks to determine the accuracy of the Fermi-Amaldi approximation. It is possible, however, that the detailed shape of $v_{\text{xc,as}}(\mathbf{r})$ is not very important as long as the energy shift in eq 18 is correct. We plan further investigations of these problems in near future.

The asymptotically corrected v_{xc} is computed on a grid and then used to obtain Kohn−Sham orbitals. These steps are repeated until convergence (with unchanged $v_{\text{xc,FA}}$, but with updated ϵ_{HOMO}). The use of asymptotically corrected v_{xc} requires modifications in Kohn−Sham procedures. In the standard Kohn−Sham DFT, v_{xc} is defined as the functional derivative of the exchange-correlation energy $E_{\text{xc}} = \int F_{\text{xc}}(\rho,\nabla\rho)d\mathbf{r}$, where F_{xc} is the exchange-correlation kernel. In generalized-gradient approximation (GGA) Kohn−Sham calculations, v_{xc} is not evaluated explicitly. Instead, integration by parts is used in all integrals involving v_{xc} to avoid second derivatives of the density. This method is not applicable to asymptotically corrected v_{xc} since F_{xc} is now not known in the asymptotic region. Therefore, we had to modify DALTON codes to be able to use the explicit formula[33] (in the short-range region)

$$v_{\text{xc}} = \frac{\partial F_{\text{xc}}}{\partial\rho} - \zeta\frac{\partial^2 F_{\text{xc}}}{\partial\zeta\partial\rho} - \frac{1}{\zeta}\frac{\partial F_{\text{xc}}}{\partial\zeta}\rho_{\gamma\gamma} - \frac{1}{\zeta^2}\left(\frac{\partial^2 F_{\text{xc}}}{\partial\zeta^2} - \frac{1}{\zeta}\frac{\partial F_{\text{xc}}}{\partial\zeta}\right)\rho_\gamma\rho_{\gamma\delta}\rho_\delta \tag{19}$$

where ρ_δ and $\rho_{\delta\gamma}$ are the first and second derivatives, respectively, of the density with respect to Cartesian coor-

dinates δ, γ, $\zeta = (\rho_x^2 + \rho_y^2 + \rho_z^2)^{1/2}$ is the length of the density gradient and explicit summation of repeated indices is assumed.

The coefficients of the converged asymptotically corrected Kohn−Sham orbitals and orbital energies are stored on disk for further processing. If a monomer-centered basis set is used, one-electron atomic integrals in a dimer basis set are computed in the next step (if a dimer-centered basis set is used, these integrals are already computed during monomer DFT calculations).

B. TD-DFT Kernel Integral. For the purpose of the TD-DFT calculations, it is necessary to evaluate, for each monomer, the matrix elements of the form[34,35]

$$\int \phi_a(\mathbf{r})\phi_r(\mathbf{r})\phi_{a'}(\mathbf{r})\phi_{r'}(\mathbf{r})\frac{\partial v_{xc}}{\partial \rho}d\mathbf{r} \tag{20}$$

where a, a' and r, r' refer to the occupied and virtual KS orbitals, respectively. Notice that in the asymptotically corrected SAPT(DFT) approach, the standard, uncorrected v_{xc} is used in eq 20 since the derivative of $v_{xc,ac}$ cannot be computed in practice. However, the orbitals are from asymptotically corrected KS calculations.

The cost of calculating eq 20 scales as $O(o^2v^2g)$, i.e, an overall $O(N^5)$ scaling, and is quite significant since the numerical integration requires very fine grids. References 20 and 22 described the implementation of the density-fitting in the evaluation of the integral 20. If the product $\phi_{a'}\phi_{r'}$ is approximated using eq 12, one only needs to compute and store the matrix elements

$$\int \phi_a(\mathbf{r})\phi_r(\mathbf{r})\chi_K(\mathbf{r})\frac{\partial v_{xc}}{\partial \rho}d\mathbf{r} \tag{21}$$

which reduces the scaling by a factor of ov/N_{aux}. The integration in eq 21 is performed using the same quadrature as in the DFT calculations.

We have implemented in ref 22 the density fitting only for the local-density approximation (LDA) kernels in eq 20 [all other stages in a TD-DFT/GGA calculation contain the proper GGA v_{xc}]. The use of the LDA kernel has been shown to result in small differences, below 1%, in dispersion energies compared to the GGA kernels.[10] This accuracy should be more than sufficient for the intended applications of the density-fitting technique, i.e., for very large systems. An implementation of this technique to GGA kernels is possible but would be significantly more complicated than in the LDA case. One should first point out that the form of integral 20 is strictly speaking valid only for the LDA v_{xc} potential, and the derivative is then just the standard partial derivative. For GGA kernels, this integral cannot be written as a product of four orbitals times a function of $\boldsymbol{r}$ independent of the indices of the orbitals. The most straightforward form of this integral [eq 25 in ref 36] includes terms containing a product of two orbitals, up to second derivatives of two other orbitals, up to second derivatives of density, and up to third derivatives of F_{xc}. This expression would be significantly more time-consuming to compute compared to the LDA kernel. If integration by parts is applied, one can obtain a more manageable integral containing only the first derivatives of density and orbital products and second derivatives of F_{xc}, see eq 23 in ref 33. Unfortunately, an application of density fitting techniques to this integral would be difficult since some terms do not contain orbital products but only their derivatives, and fitting such derivatives is numerically more difficult than fitting orbital products only.

C. Integral Transformation. All quantities in eq 8 are given in terms of one- and two-electron integrals over molecular orbitals (and propagators for the CKS terms). After the techniques presented in refs 10, 20, and 22 (see also section IV D) are applied and the $O(N^6)$ terms eliminated, the two-electron transformation with the nominal scaling of $O(N^5)$ becomes the most time-consuming part of SAPT(DFT) for a wide range of systems. Although there are other components scaling as $O(N^5)$ remaining, the conventional transformation has the largest prefactor. However, as it will be shown below, the transformation can greatly benefit from density fitting.

The objective of the two-electron transformation is to compute molecular integrals

$$(ij|kl) = \int d\mathbf{r}_1 d\mathbf{r}_2 \frac{\phi_i(\mathbf{r}_1)\phi_j(\mathbf{r}_1)\phi_k(\mathbf{r}_2)\phi_l(\mathbf{r}_2)}{|\mathbf{r}_1 - \mathbf{r}_2|} \tag{22}$$

where i, j, k, and l label molecular (occupied or virtual) orbitals of monomer A or B, expanded in terms of atomic (and possibly midbond) basis functions ψ_μ, e.g., $\phi_p = \sum_\mu c_{\mu p}\psi_\mu$. The integral written above can be expressed using the densities of eq 11 as

$$(ij|kl) = \int d\mathbf{r}_1 d\mathbf{r}_2 \frac{\rho_{ij}(\mathbf{r}_1)\rho_{kl}(\mathbf{r}_2)}{|\mathbf{r}_1 - \mathbf{r}_2|} \tag{23}$$

By inserting eq 12 into eq 23, one obtains

$$(ij|kl) \approx \sum_{KL} D_K^{ij} D_L^{kl} J_{KL} = \sum_K D_K^{ij} \tilde{D}_K^{kl} \tag{24}$$

where

$$\tilde{D}^{ij}{}_K = \sum_L D_{KL}^{ij} J_L \tag{25}$$

with J_{KL} defined in eq 16. Since we use dimer-centered auxiliary basis sets, χ_K and χ_L are always from the same basis set. By inserting eq 14 into eq 25, we obtain

$$\tilde{D}_K^{kl} = (kl|K) \tag{26}$$

and eq 24 then reads

$$(ij|kl) \approx \sum_K D_K^{ij}(kl|K) \tag{27}$$

The integrals $(ij|K)$ are obtained by transforming the 3-center atomic orbital (AO) integrals $(\mu\nu|K)$

$$(\mu j|K) = \sum_\nu c_{\nu j}(\mu\nu|K) \tag{28}$$

$$(ij|K) = \sum_\mu c_{\mu i}(\mu j|K) \tag{29}$$

where c_{vi} are molecular (Kohn−Sham) orbital coefficients and $(\mu i|K)$ are partially transformed integrals.

In our implementation, the 3-center $(\mu\nu|K)$ and 2-center J_{KL} integrals are first calculated in the full, dimer-centered basis set using the integral package adopted from the GAMESS-US code.[37] If monomer-centered basis sets are used, the required integrals for monomer A and B are also extracted from this file. Then, the 3-center integrals are contracted according to eq 28, where the orbital coefficients may correspond to the monomer A or B depending on the computed integral. This step scales as $O(wn^2N_{\mathrm{aux}})$, where $w = o$ or v depending on the contraction index i, and n is the total number of orbitals $n = o + v$. One cannot avoid i being a virtual index for $(ab|rs)$-type integrals only, where a, b denote occupied and r, s denote virtual indices (these integrals are required for the TD-DFT matrices with hybrid functionals and for the third-order terms, see section IV D). Since in this case the resulting set of semitransformed integrals $(\mu j|K)$ can be large, the transformation is done out-of-core, i.e., with only a part of the integral matrix stored in memory. Although such a transformation requires more disk operations than an in-core one, it scales only as $O(N^4)$, and, therefore, for larger systems it is only a small part of the whole calculation. The intermediate semitransformed integrals are stored and reused for all integrals that share the same intermediate. Next, the contractions of eq 29 are performed, and the resulting 3-center integrals are also stored. This step scales as $O(ww'N_{\mathrm{aux}})$, where $w' = o$ or v depending on the contraction index j. The smaller sets of the 3-index intermediates are stored in memory, whereas the larger ones are stored on disk and then read in batches that can fit into memory.

After the integrals are computed, the matrix $\mathbf{J}$ of eq 16 is inverted in an $O(N_{\mathrm{aux}}^3)$ step, and then the fit coefficients of eq 14 are computed at a cost of $O(ww'N_{\mathrm{aux}}^2)$ (with w, w' equal to o or v) and stored on disk as the file size is $ww'N_{\mathrm{aux}}$ which can be fairly large. Finally, the transformed 4-index integrals are obtained from eq 27 and stored on disk. This final step scales at most as $O(o^2v^2N_{\mathrm{aux}})$, or $O(N^5)$, the highest formal scaling in the whole transformation procedure. The o^2v^2 part of the scaling is due to the fact that molecular integrals required by SAPT(DFT) have no more than two virtual indices.

Although the transformation described above has the same nominal $O(N^5)$ scaling as the conventional transformation, the operation count is reduced compared to the latter one [the cost of the latter procedure is $O(on^4)$]. Moreover, the density fitting brings additional advantages[19] due to the smaller size of the atomic integral file (3-index, instead of 4-index), reducing the IO-operation count and simplifying the usage of very efficient matrix algebra routines in eq 24.

D. CKS Induction Energies. We have shown in ref 22 how the CKS dispersion energies can be efficiently computed using density-fitting techniques. Here we apply the same method to the CKS induction energies. The CKS FDDS appearing in eqs 5 and 6 can be expressed[35,38] as linear combinations of products of occupied and virtual orbitals of a given monomer. For monomer A we have

$$\alpha_A(\boldsymbol{r},\boldsymbol{r}'|iu) = \sum_{ara'r'} C^A_{ar,a'r'}(iu)\phi_a(\boldsymbol{r})\phi_r(\boldsymbol{r})\phi_{a'}(\boldsymbol{r}')\phi_{r'}(\boldsymbol{r}') \quad (30)$$

where the coefficient matrix $\mathbf{C}^{\mathrm{A}}(iu)$ is obtained from the equation

$$(\mathbf{H}^{(2)}\mathbf{H}^{(1)} + u^2\mathbf{I}_{ov})\mathbf{C}^A(iu) = -4\mathbf{H}^{(2)} \quad (31)$$

with $\mathbf{I}_{ov}$ denoting the $ov \times ov$ unit matrix and the matrices $\mathbf{H}^{(i)}$, $i = 1, 2$, given by[34]

$$\mathbf{H}^{(1)} = \mathbf{d} + 4\mathbf{H}_0^{(1)} + \mathbf{H}_{\mathrm{r}}^{(1)} \quad (32)$$

$$\mathbf{H}^{(2)} = \mathbf{d} + \mathbf{H}_{\mathrm{r}}^{(2)} \quad (33)$$

The diagonal matrix $\mathbf{d}$ is defined in terms of orbital energies ϵ_p as $\mathbf{d}_{ar,a'r'} = \delta_{aa'}\delta_{rr'}(\epsilon_r - \epsilon_a)$, whereas the remaining matrices are given by[34]

$$(\mathbf{H}_0^{(1)})_{ar,a'r'} = (ar|a'r') + \int \phi_a\phi_r\phi_{a'}\phi_{r'}\frac{\partial v_{\mathrm{xc}}}{\partial\rho}d\mathbf{r} \quad (34)$$

$$(\mathbf{H}_{\mathrm{r}}^{(1)})_{ar,a'r'} = -\xi[(aa'|rr') + (ar'|a'r)] \quad (35)$$

$$(\mathbf{H}_{\mathrm{r}}^{(2)})_{ar,a'r'} = -\xi[(aa'|rr') - (ar'|a'r)] \quad (36)$$

where $0 \le \xi \le 1$ is the fraction of the Hartree−Fock exchange applied in a given DFT functional. Notice that although we use LDA kernel in eq 34 in the density-fitted approach, the parameter ξ is the same as in the hybrid DFT method applied. The exchange part of the LDA's v_{xc} is appropriately scaled. Some terms representing current density, which give negligible contributions for the problems considered here,[35] have been neglected in $\mathbf{H}^{(2)}$. Equations analogous to eqs 30−36—with orbitals and orbital energies of monomer A replaced by those of monomer B—describe the propagator $\alpha_B(\boldsymbol{r},\boldsymbol{r}'|iu)$.

To place calculations of the CKS induction energies in the context of the complete SAPT(DFT) calculations, let us briefly recall the approach of refs 8, 10, and 22 for the CKS dispersion energies. Inserting expansions 30 for monomers A and B into formula 5, the following expression for the CKS dispersion energy is obtained

$$E_{\mathrm{disp}}^{(2)}(\mathrm{CKS}) = -\frac{1}{2\pi}\int_0^\infty du \sum_{ara'r'bsb's'} C^A_{ar,a'r'}(iu)\, C^B_{bsb's'}(iu)(ar|bs),(a'r'|b's') \quad (37)$$

The time requirements of expression 37 and the solution of eq 31 both scale as $O(o^3v^3)$ or $O(N^6)$. Thus, for large o and v, evaluation of $E_{\mathrm{disp}}^{(2)}$(CKS) from eq 37 becomes the most time-consuming step of the SAPT(DFT) calculation. It should be emphasized at this point that the calculation of dispersion energy according to formula 37 requires the *full* propagator matrices $\mathbf{C}^A(iu)$ and $\mathbf{C}^B(iu)$ computed at a number of imaginary frequencies. This is in contrast to typical TD-DFT calculations, where only a *vector* quantity $\mathbf{C}^{\mathrm{X}}\mathbf{w}$ is of interest, where $\mathbf{w}$ is a column vector of ov matrix elements representing some perturbation of the system. Multiplying both sides of eq 31 on the right by $\mathbf{w}$, one obtains a system of linear

equations for $\mathbf{C}^X\mathbf{w}$ with the vector $-4\mathbf{H}^{(2)}\mathbf{w}$ as the right-hand side. Such a system can then be solved using iterative techniques[39,40] involving only matrix-vector multiplications and scaling as $(ov)^2$ or $O(N^4)$. A procedure of this type will be employed by us to compute the CKS induction energy. The scaling of expression 37 can be reduced to $O(N^3)$ if the two-electron integrals are approximated with density fitting.[8,10] One then obtains

$$E_{\mathrm{disp}}^{(2)} = -\frac{1}{2\pi}\int_0^\infty du \sum_{KL}^{N_{\mathrm{aux}}^A}\sum_{K'L'}^{N_{\mathrm{aux}}^B} \tilde{C}_{KL}^A(iu)\,\tilde{C}_{K'L'}^B(iu) J_{KK'}J_{LL'} \quad (38)$$

where the $N_{\mathrm{aux}}^X \times N_{\mathrm{aux}}^X$ matrix $\tilde{\mathbf{C}}^X$, $X = A, B$, is the result of the transformation

$$\tilde{\mathbf{C}}^X \equiv (\mathbf{D})^t\mathbf{C}^X\mathbf{D} \quad (39)$$

with $\mathbf{D}$ being the $ov \times N_{\mathrm{aux}}^X$ density-fitting coefficient matrix of monomer X, given by eq 14 (for simplicity of discussion we will further assume that $N_{\mathrm{aux}}^A = N_{\mathrm{aux}}^B = N_{\mathrm{aux}}$). Although the cost of performing transformation 39 is $O(o^2v^2N_{\mathrm{aux}})$ and that of evaluating expression 38 is only $O(N_{\mathrm{aux}}^3)$, the overall scaling of dispersion energy is still $O(N^6)$, due to $O(o^3v^3)$ cost of matrix operations necessary to solve eq 31 for $\mathbf{C}^X$. In ref 22 it has been shown that the latter step can be bypassed, and $\tilde{\mathbf{C}}^X$ can be obtained directly from a fast-converging iterative procedure scaling as $O(o^2v^2N_{\mathrm{aux}})$, or $O(N^5)$, in the case of hybrid functionals ($\xi \neq 0$). For nonhybrid functionals ($\xi = 0$), the iterative algorithm reduces to a one-step procedure, identical to the one described in ref 20, scaling as $O(ovN_{\mathrm{aux}}^2)$, or $O(N^4)$. Thus, using density fitting and the techniques of ref 22, the $E_{\mathrm{disp}}^{(2)}$(CKS) energy is obtained at a cost of at most $O(N^5)$. It has been shown[10,20,22] that the errors in this quantity resulting from density fitting approximation and truncation of the iterative scheme are negligible even if $N_{\mathrm{aux}} \ll ov$.

Returning now to the CKS induction energy, $E_{\mathrm{ind}}^{(2)}$(CKS), the first term on the right hand side of eq 6 can be rewritten using the orbital representation 30 of the CKS propagator $\alpha_A(\mathbf{r}_1, \mathbf{r}_1'|0)$

$$E_{\mathrm{ind}}^{(2)}(A \leftarrow B) = \frac{1}{2}\int \omega_B(\mathbf{r})\alpha_A(\mathbf{r},\mathbf{r}'|0)\omega_B(\mathbf{r}')d\mathbf{r}d\mathbf{r}' = \frac{1}{2}\boldsymbol{\omega}^t\mathbf{C}^A(0)\boldsymbol{\omega} = \frac{1}{2}\boldsymbol{\omega}^t\mathbf{Z} \quad (40)$$

where we used the definition $\mathbf{Z} \equiv \mathbf{C}^A(0)\boldsymbol{\omega}$ and $\mathbf{C}^A(0)$ is obtained by solving eq 31 at zero frequency. The ov elements of the vector $\boldsymbol{\omega}$ are given by

$$\omega_{ar} = \int \phi_a(\mathbf{r})\phi_r(\mathbf{r})\omega_B(\mathbf{r})d\mathbf{r} \quad (41)$$

(Analogous expressions for $E_{\mathrm{ind}}^{(2)}(B \leftarrow A)$ can be obtained by properly exchanging monomer indices). Setting $u = 0$ in eq 31, then multiplying both sides of this equation on the right by ω and on the left by $(\mathbf{H}^{(2)})^{-1}$, one finds that $\mathbf{Z}$ satisfies the equation

$$\mathbf{H}^{(1)}\mathbf{Z} = -4\boldsymbol{\omega} \quad (42)$$

The matrix $\mathbf{H}_0^{(1)}$, eq 34, involving the four-index integrals considered in section IV B, can be written using eqs 12 and 21 as $\mathbf{FD}^t$, where the elements of the $ov \times N_{\mathrm{aux}}$ matrix $\mathbf{F}$ are defined as

$$F_{ar,K} = (ar|K) + \int \phi_a(\mathbf{r})\phi_r(\mathbf{r})\chi_K(\mathbf{r})\frac{\partial v_{\mathrm{xc}}}{\partial \rho}d\mathbf{r} \quad (43)$$

Equation 42 then becomes

$$(\mathbf{d} + 4\mathbf{FD}^t + \mathbf{H}_{\mathrm{r}}^{(1)})\mathbf{Z} = -4\boldsymbol{\omega} \quad (44)$$

Although eq 44 is simpler than its equivalent in the case of dispersion energy calculations [cf. eq 14 of ref 22], its direct solution scales also as $(ov)^3$. However, the matrix $\boldsymbol{\Lambda} \equiv \mathbf{d} + 4\mathbf{FD}^t$ can be inverted at a much lower cost. This can be achieved by using eq 16 in ref 22 to write the inverse of $\boldsymbol{\Lambda}$ as

$$\boldsymbol{\Lambda}^{-1} = \mathbf{d}^{-1} - 4\mathbf{d}^{-1}\mathbf{F}(\tilde{\mathbf{I}} + 4\mathbf{D}^t\mathbf{d}^{-1}\mathbf{F})^{-1}\mathbf{D}^t\mathbf{d}^{-1} \quad (45)$$

where $\tilde{\mathbf{I}}$ is the $N_{\mathrm{aux}} \times N_{\mathrm{aux}}$ unit matrix. The operations involved in eq 45 are performed in the following way. First, the matrix in parentheses is constructed from matrices $\mathbf{F}$ and $\mathbf{D}$ stored in memory, which requires matrix−matrix multiplications scaling as N_{aux}^2ov, or, equivalently, $O(N^4)$. The inverse of this matrix is then obtained at a cost proportional to N_{aux}^3 and stored in memory. The remaining matrix multiplications in eq 45 could be performed at the cost $O(N^4)$. However, matrix $\boldsymbol{\Lambda}^{-1}$ is never computed explicitly. Instead, its action on a vector of length ov is evaluated as a sequence of matrix-vector multiplications using consecutive matrices in eq 45, with scaling not exceeding $N_{\mathrm{aux}}ov$, or $O(N^3)$.

Application of $\boldsymbol{\Lambda}^{-1}$ to both sides of eq 44 written as $\boldsymbol{\Lambda}\mathbf{Z} = -4\boldsymbol{\omega} - \mathbf{H}_{\mathrm{r}}^{(1)}\mathbf{Z}$ leads to an iterative process

$$\mathbf{Z}_{n+1} = \mathbf{Z}_0 - \boldsymbol{\Lambda}^{-1}\mathbf{H}_{\mathrm{r}}^{(1)}\mathbf{Z}_n \quad (46)$$

with $\mathbf{Z}_0 = -4\boldsymbol{\Lambda}^{-1}\omega$. The iterations stop when the length of dimensionless vector $\mathbf{Z}$ changes by less than a predefined threshold, set equal to 10^{-12}. For nonhybrid functionals, i.e., when $\mathbf{H}_{\mathrm{r}}^{(1)} = 0$, the solution $\mathbf{Z} = \mathbf{Z}_0$ is obtained in a one-step procedure. The term $\boldsymbol{\Lambda}^{-1}\mathbf{H}_{\mathrm{r}}^{(1)}\mathbf{Z}_n$ is computed in each iteration by first evaluating the vector $\mathbf{H}_{\mathrm{r}}^{(1)}\mathbf{Z}_n$ and then multiplying this vector by $\boldsymbol{\Lambda}^{-1}$ in the way described above. The former of these steps, scaling as $(ov)^2$ or $O(N^4)$, is the most demanding part of the whole calculation of the induction energy. Still, this scaling is much more favorable than the $O(N^5)$ requirements of several other steps in a SAPT(DFT) calculation. It should be also mentioned that there is no need to store the entire matrix $\mathbf{H}_{\mathrm{r}}^{(1)}$ in memory in order to perform a matrix-vector multiplication. Instead, a number of rows of this matrix is read in at a time, depending on available memory, and the corresponding components of the resultant vector $\mathbf{H}_{\mathrm{r}}^{(1)}\mathbf{Z}_n$ are evaluated. This is important for larger systems, where the $(ov)^2$ storage requirement would be sizable.

E. SAPT(KS) Terms. All terms $E_{\mathrm{a}}^{(n)}$(KS) in eqs 8 and 10, where 'a' stands for any of the electrostatic, induction, dispersion, or exchange terms, are computed from the standard SAPT expressions for $E_{\mathrm{a}}^{(n0)}$ in which the Hartree−

Fock orbitals and orbital energies have been replaced by their Kohn−Sham counterparts. The relevant expressions, presented in ref 41, are evaluated using the existing routines from the SAPT2002 program.[23] The required one- and two-electron integrals over KS molecular orbitals are obtained from integrals over atomic orbitals as described in section IV C. The calculation of the KS terms from molecular orbitals is quite fast, even though, as discussed earlier, the most time-consuming of these terms, $E^{(2)}_{\mathrm{exch-disp}}$(KS), scales as $O(o^3v^2)$, or $O(N^5)$. However, since normally $o \ll v$, evaluation of this term is much faster than that of the other $O(N^5)$ parts of a SAPT(DFT) calculation.

F. Higher Orders in *V*. In some cases, the effects of higher orders in V are significant and have to be included. In past applications of SAPT, these effects have been usually estimated as the difference between the Hartree−Fock interaction energy and the sum of SAPT terms up to the second order in V that do not include any correlation effects

$$\delta E^{\mathrm{HF}}_{\mathrm{int}} = E^{\mathrm{HF}}_{\mathrm{int}} - (E^{(10)}_{\mathrm{elst}} + E^{(10)}_{\mathrm{exch}} + E^{(20)}_{\mathrm{ind,resp}} + E^{(20)}_{\mathrm{exch-ind,resp}}) \quad (47)$$

where the quantities with the subscript "resp" are computed including the coupled Hartree−Fock-type response of monomer orbitals to the field of the partner. The quantity $\delta E^{\mathrm{HF}}_{\mathrm{int}}$ is a good approximation to higher-order terms in case of molecules with a significant induction contribution. For molecules with a small induction contribution, the benefits of including $\delta E^{\mathrm{HF}}_{\mathrm{int}}$ are not clear, and, in some cases, like rare gas dimers, this component does not approximate the third and higher-order effects well. Although including $\delta E^{\mathrm{HF}}_{\mathrm{int}}$ does not increase the scaling beyond that of SAPT(DFT), in most cases the supermolecular SCF, together with the terms of regular SAPT listed in eq 47, form a significant part of the whole calculation.

Recently, explicit formulas for the third-order terms have been derived and implemented.[42] The sum of the induction and exchange-induction terms, $E^{(30)}_{\mathrm{ind}} + E^{(30)}_{\mathrm{exch-ind}}$, can provide a major part of high-order effects. Preliminary tests with wave function-based SAPT[42] showed that this approach leads to more accurate interaction energies for nonpolar systems than the use of $\delta E^{\mathrm{HF}}_{\mathrm{int}}$. The two corrections can be straightforwardly computed in the SAPT(KS) approach. The density-fitting formalism has been applied to obtain the molecular integrals needed [with $O(o^2v^2N_{\mathrm{aux}})$ scaling]. The use of $E^{(30)}_{\mathrm{ind}}$ and $E^{(30)}_{\mathrm{exch-ind}}$ does not increase the overall scaling of SAPT(DFT) as the cost of these corrections scales as $O(o^2v^2)$ and $O(o^3v^2)$, respectively. Since the third-order terms in the KS version have not yet been sufficiently tested, we have not included them in the numerical results presented below.

G. Advantages of Density Fitting. Concluding this section, let us shortly summarize the advantages of using density-fitting with SAPT(DFT). Compared to the approach without density fitting, the method gains an order of magnitude better scaling. The cost of the dispersion energy calculation reduces from $O(o^3v^3)$ to $O(o^2v^2N_{\mathrm{aux}})$, and the computation of the matrix elements involving the exchange-correlation kernel requires $O(ovN_{\mathrm{aux}}g)$ operations instead of $O(o^2v^2g)$ needed in the standard case. Scaling of the most expensive step of the transformation is reduced from on^4 to $o^2v^2N_{\mathrm{aux}}$. Memory requirements of transformation and of the CKS-based calculations are also significantly reduced since most operations are performed on 3-index objects which fit in memory easier than the 4-index ones used in the standard SAPT(DFT). With reduced memory usage, it is straightforward to apply highly optimized matrix−matrix multiplication BLAS routines,[43] which results in further speedups. Since no 4-index AO integrals are needed, only 3-index and relatively small (o^2v^2) 4-index objects have to be stored on disk. This results in a very significant reduction of disk usage and the input/output (I/O) operation count.

V. Results and Discussion

A. Numerical Details. We have tested the density-fitting SAPT(DFT) approach mainly on the example of the benzene dimer for which we have considered three intermolecular separations. Additional tests have been performed for near-equilibria configurations of the argon dimer, the water dimer, and the dimer of cyclotrimethylene trinitramine, $(CH_2-N-NO_2)_3$, known also under the name RDX. The Kohn−Sham orbitals were obtained using the PBE0 functional[44,45] with the Fermi-Amaldi asymptotic correction and Tozer-Handy splicing scheme[31,46] computed with the experimental ionization potentials[47] equal to 0.3397, 0.5791, and 0.4638 hartree for C_6H_6, Ar, and H_2O, respectively, and 0.373 hartree for RDX computed using PBE0 as the energy difference between the neutral molecule and the cation. In all calculations the LDA kernel was used in eqs 20 and 21. The DFT calculations were performed using the DALTON code[29] with the aug-cc-pV*X* Z, $X = 2$, 3 and cc-pVDZ bases of Dunning et al.[48] In some cases, these basis sets were centered on both monomers and extended with a set of midbond functions, placed halfway between the centers of mass of the monomers. A dimer-centered basis set (DCBS) containing midbond functions will be referred to as DC$^+$BS. In another approach, referred to as MC$^+$BS[24], orbitals of a given monomer were expanded in terms of this monomer's own basis, the midbond functions, and the isotropic part (i.e., with the polarization functions removed) of the basis of the other monomer. Our set of midbond functions (3*s*3*p*2*d*2*f*) consisted of three *s* and three *p* shells with exponents (0.9, 0.3, 0.1) and of two *d* and two *f* shells with exponents (0.6, 0.2). The density fitting approximation was accomplished for all SAPT terms using auxiliary basis sets of ref 49, fitted to the second-order Møller−Plesset (MP2) results for atoms and corresponding to the principal orbital bases applied. As it was the case with the underlying principal bases, the auxiliary ones were usually extended with a set of midbond functions, containing five each of uncontracted *spd* shells with exponents (1.8, 1.2, 0.6, 0.4, 0.2), four *f* shells with exponents (1.5, 0.9, 0.5, 0.3), and three *g* shells with exponents (1.5, 0.9, 0.3), chosen to approximately reproduce the products of midbond functions. Only the DCBS and DC$^+$BS types (but not MC$^+$BS) were used for auxiliary bases (even if the principal basis set was of MC$^+$BS type). Although the MC$^+$BS auxiliary functions would reduce slightly the computational cost of some parts of the code, the resulting inability of reusing certain intermediates during the transformation and a small loss of accuracy would outweigh the

Table 1: Decomposition of the SAPT(DFT) Interaction Energy Obtained with Density Fitting for the Benzene Dimer in a "Sandwich" Configuration with Monomers in the Geometry of Ref 50[a]

	$R = 3.2$ Å		$R = 3.85$ Å		$R = 5.0$ Å	
$E_{elst}^{(1)}$(KS)	−6.1931	(−0.0011)	0.1362	(0.0019)	0.5550	(−0.0009)
$E_{exch}^{(1)}$(KS)	21.8896	(0.0021)	3.2976	(0.0006)	0.0944	(0.0001)
$E_{ind}^{(2)}$(CKS)	−8.7913	(0.0002)	−1.1067	(0.0000)	−0.0626	(−0.0000)
$\tilde{E}_{exch-ind}^{(2)}$(CKS)	8.4937	(−0.0002)	0.8947	(−0.0000)	0.0121	(0.0000)
$E_{disp}^{(2)}$(CKS)	−15.0480	(0.0037)	−5.3452	(0.0008)	−1.1005	(−0.0000)
$\tilde{E}_{exch-disp}^{(2)}$(CKS)	2.3979	(−0.0009)	0.4480	(−0.0002)	0.0179	(−0.0000)
$E_{int}^{SAPT(DFT)}$	2.7489	(0.0038)	−1.6753	(0.0030)	−0.4838	(−0.0008)

[a] The aug-cc-pVDZ MC$^+$BS basis set with the 3*s*3*p*2*d*2*f* midbond set was used. The errors resulting from density fitting are given in parentheses. The unit for the energies and the errors is kcal/mol. All calculations in double precision.

benefits. The frequency integral in formula 5 for the dispersion energy was evaluated using an 8-point Gauss-Legendre quadrature, and the first two terms were used in expansion 27 of ref 22 (i.e., two iterations were performed in solving the TD-DFT set of equations for the propagator matrix $\tilde{\mathbf{C}}$).

B. Benzene Dimer. The results for the benzene dimer in the parallel ("sandwich") geometry are presented in Table 1. The intermolecular distances were chosen to range from 3.2 Å (repulsion wall), through 3.85 Å (minimum), to 5.0 Å (long-range region). As the table shows, for all the distances the accuracy of density fitting is very satisfactory, the error always being well below 0.01 kcal/mol for all interaction components. For the total interaction energies, the largest discrepancy, 0.018% at the minimum geometry, is much smaller than the error resulting from the incompleteness of the basis set (cf. the results for the dispersion energy with the aug-cc-pVTZ in ref 22). Although the calculation of dispersion energy involves an additional approximation besides density fitting of integrals, namely the truncation of the expansion 27 in ref 22, the error of this component does not dominate the total error, except for the small distances, but even then the error is very small. The relative error is largest for electrostatic term, exceeding 1% for the minimum geometry. This component will be discussed in the next subsection.

C. Accuracy of the Electrostatic Component. As discussed above, the relative errors of density fitting are usually the largest for the electrostatic energy. It is easy to understand why the electrostatic term is difficult to fit. This component, obtained by summing the positive contributions of the electron−electron and nuclear−nuclear repulsion interactions with the negative electron−nuclear attraction term, is typically much smaller in magnitude than either of these three terms. The error introduced by density fitting, which affects only the electron repulsion term, is, in fact, very small, amounting to just 1.4×10^{-6}% of this term for benzene dimer at 3.85 Å. However, this error may still become comparable to the total electrostatic energy, which makes the latter correction particularly sensitive to the quality of the fit.

One way to improve the accuracy of the electrostatic energy is to use larger and/or better auxiliary basis sets. In ref 20, Hesselmann et al. used two different types of auxiliary bases. The so-called JK-optimized bases (named for the symbols denoting the Coulomb and exchange integrals) of ref 51 were used for all SAPT components except for the dispersion and exchange-dispersion energies. For the two latter components, the MP2-optimized bases of ref 49 were applied. The JK auxiliary bases are better suited than the MP2-optimized ones to reproduce products of occupied orbitals requiring large exponents. Thus, in particular the electrostatic and first-order exchange energies may be better fitted by JK bases since these terms depend only on occupied orbitals. Since no JK-basis sets corresponding to the augmented bases of Dunning et al. are available, the authors of ref 20 suggested to use the JK auxiliary bases optimized for the cc-pV(X+1) basis sets, i.e., to increment the cardinal number by one relative to the principal basis set used. When we applied JK basis sets for the benzene dimer, we found that this resulted in some significant numerical instabilities in the electrostatic term. In particular, the results differed by about 0.001 kcal/mol between different computer architectures. It turned out that these problems were not due to the use of the JK-type bases but to the size of the basis sets leading to linear dependencies. For example, the use of the MP2-optimized aug-cc-pVTZ auxiliary basis for the benzene dimer resulted in differences between architectures up to 0.03 kcal/mol (whereas the aug-cc-pVDZ results reported in Table 1 are stable). We have found that the main sources of this numerical error were the inversion of the **J** matrix and the summation of eq 14. By performing these two calculations in quadruple precision, this numerical error can be reduced by several orders of magnitude. We have also tested the singular value decomposition (SVD) method recommended in ref 52 for such cases. With 10^{-7} threshold for neglecting small singular values, the numerical stability was improved, but the overall density-fitting error increased. Therefore, it appears that the use of quadruple precision performs better. Since the inversion of the **J** matrix scales as $O(N^3)$ and is done only once for the whole SAPT(DFT) calculation and the calculation of the **D** matrix of eq 14 for the electrostatic term scales as $O(o^2N_{aux}^2)$, even with quadruple precision both calculations are a small fraction of the total costs. We recommend the quadruple precision approach for auxiliary basis sets larger than about 1400 functions since the numerical precision error becomes larger at this size than the density-fitting error. For all other SAPT(DFT) terms, the numerical instability is below 0.0001 kcal/mol, even for the largest auxiliary basis sets tested, and the quadruple precision is not necessary.

With the quadruple precision procedure, we found that the JK-optimized basis sets give more accurate results for the

Table 2: Decomposition of the SAPT(DFT) Interaction Energy Obtained with Density Fitting for Two Argon Atoms Separated by 3.75 Å[a]

	aug-cc-pVDZ		aug-cc-pVTZ	
$E^{(1)}_{elst}$(KS)	−51.451	(−0.446)	−49.718	(0.364)
$E^{(1)}_{exch}$(KS)	169.504	(0.022)	169.642	(0.002)
$E^{(2)}_{ind}$(CKS)	−66.910	(0.000)	−65.850	(−0.000)
$\tilde{E}^{(2)}_{exch-ind}$(CKS)	65.582	(−0.003)	64.718	(0.000)
$E^{(2)}_{disp}$(CKS)	−222.617	(−1.647)	−229.297	(−0.002)
$\tilde{E}^{(2)}_{exch-disp}$(CKS)	15.735	(−0.089)	16.227	(−0.044)
$E^{SAPT(DFT)}_{int}$	−90.158	(−2.163)	−94.280	(0.319)

[a] The aug-cc-pVDZ and aug-cc-pVTZ DC$^+$BS basis sets with the 3*s*3*p*2*d*2*f* midbond set were used. The errors resulting from density fitting are given in parentheses. The unit for the energies and the errors is cm^{-1}. All calculations in double precision.

electrostatic component than the MP2-optimized auxiliary basis of similar size. For benzene dimer at 3.85 Å and the aug-cc-pVDZ principal basis set, the cc-pVTZ JK auxiliary basis set (N_{aux} = 1408) gives 0.0003 kcal/mol density-fitting error in comparison to the 0.0019 kcal/mol error for the MP2-optimized aug-cc-pVDZ auxiliary basis set (N_{aux} = 1240). For the aug-cc-pVTZ principal basis set, the density-fitting errors are −0.002 kcal/mol and −0.005 kcal/mol for JK-cc-pVTZ and MP2-aug-cc-pVTZ (N_{aux} = 1924) auxiliary bases, respectively. Thus, if a very high accuracy of the electrostatic component is required, we recommend the use of JK-optimized bases for this component. In most cases, however, the MP2-optimized bases should be adequate for all components.

D. Argon Dimer. In Table 2, we present SAPT(DFT) results for the argon dimer at the van der Waals minimum. For the aug-cc-pVDZ case, the largest fitting error is in the dispersion term, and the resulting total energy error is 2.5%. This error is considerably larger than for other examples tested. Thus, the argon auxiliary basis set of ref 49 appears somewhat less accurate than other bases of the same size. Still, the error resulting from density fitting is a few times smaller than the basis set incompleteness error which is about 8% (cf. results in the aug-cc-pVQZ basis in ref 13). Therefore, the aug-cc-pVDZ auxiliary basis is in fact adequate. However, the aug-cc-pVTZ basis set performs much better. The largest fitting error is now in the electrostatic term, and the total energy error is only about 0.3%.

E. Water Dimer. Water dimer has been chosen as an example of a polar system. The results presented in Table 3 show that the accuracy of density-fitting is excellent with the largest percentage error of 0.01% in the exchange-dispersion energy. Although in this case the density-fitting errors are negligible already when the aug-cc-pVDZ basis set is used, the aug-cc-pVTZ basis reduces these errors further, as for the argon dimer. This is a very positive observation since with larger basis sets one aims for higher accuracy. Notice that for the water dimer as well as for the benzene and argon dimers the induction and exchange-induction components always exhibit the highest accuracy, exceeding in most cases the number of significant figures presented in the tables. Apparently, the polarization phenomenon results in smooth, easy to fit densities.

Table 3: Decomposition of the SAPT(DFT) Interaction Energy Obtained with Density Fitting for the Water Dimer in a Geometry Close to the Global Minimum[a]

	aug-cc-pVDZ		aug-cc-pVTZ	
$E^{(1)}_{elst}$(KS)	−6.8900	(0.0014)	−6.8847	(−0.0002)
$E^{(1)}_{exch}$(KS)	5.7418	(0.0027)	5.7399	(0.0003)
$E^{(2)}_{ind}$(CKS)	−2.4185	(0.0007)	−2.5381	(0.0000)
$\tilde{E}^{(2)}_{exch-ind}$(CKS)	1.2461	(−0.0001)	1.3417	(0.0000)
$E^{(2)}_{disp}$(CKS)	−2.0350	(0.0001)	−2.3806	(0.0004)
$\tilde{E}^{(2)}_{exch-disp}$(CKS)	0.3420	(−0.0008)	0.4037	(−0.0005)
$E^{SAPT(DFT)}_{int}$	−4.0136	(0.0040)	−4.3180	(−0.0000)

[a] Geometry as in ref 11 with R = 3 Å. The aug-cc-pVDZ and aug-cc-pVTZ basis sets in the DCBS form were used (without midbond). The errors resulting from density fitting are given in parentheses. The unit for the energies and the errors is kcal/mol. All calculations in double precision.

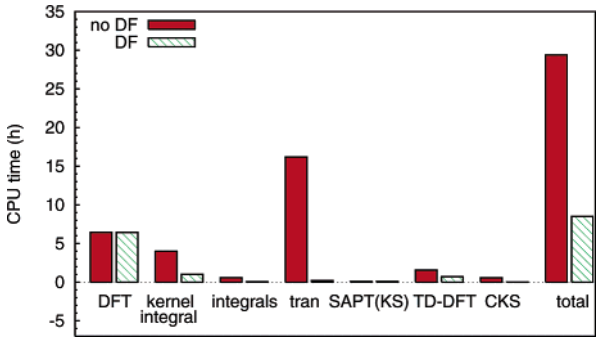


Figure 1. Wall times for the benzene dimer on the 2.4 GHz Opteron processor. The aug-cc-pVDZ basis set with 3*s*3*p*2*d*2*f* midbond was used, corresponding to o = 21, v = 303, N_{aux} = 1240, g = 389 448 grid points in the DFT and TD-DFT calculations. 'DF'—timings with density fitting. 'no DF'—standard DALTON-based SAPT(DFT) with LDA kernel. 'DFT'—two monomer DFT calculations; 'kernel integral'—eq 21 for the DF approach or eq 20 for the no-DF approach for both monomers; 'integrals'—3-index (DF) or 4-index (no-DF) integrals of the dimer; 'tran'—integral transformation; 'SAPT(KS)'—total time for SAPT(KS) terms; 'TD-DFT'——time of computing TD-DFT propagators, 'CKS'—the CKS induction and dispersion energies.

F. Timings of SAPT(DFT). Figure 1 shows the wall-clock times of various steps in the calculations for the benzene dimer, using both the standard and density-fitted SAPT(DFT) approaches. Overall, density fitting accelerates this calculation more than three times with much larger speedups for some of the components. While the major speedup occurs in the transformation step, substantial improvements are also visible in the timings of the calculation of the integrals of eqs 20 and 21 and of the TD-DFT and CKS calculations. For systems such as those treated in this work, the overall timing of the density-fitting approach is now dominated by the monomer Kohn−Sham calculations. This is partly due to the fact that the Kohn−Sham code used by us does not yet take advantage of density fitting (this is also the reason that the overall speedup is only a factor of 3). It is also worth pointing out that a supermolecular DFT calculations for this system would take more time than the SAPT(DFT) calculation (and would produce a worthless result). For larger systems, the benefit of density fitting will show up mostly

Table 4: Decomposition of the SAPT(DFT) Interaction Energy Obtained with Density Fitting for RDX Dimer in a Geometry Extracted from Crystal Structure[53] and Specified in the Supporting Information[54] [a]

$E^{(1)}_{elst}$(KS)	−4.984
$E^{(1)}_{exch}$(KS)	2.867
$E^{(2)}_{ind}$(CKS)	−1.080
$\tilde{E}^{(2)}_{exch-ind}$(CKS)	0.465
$E^{(2)}_{disp}$(CKS)	−3.487
$\tilde{E}^{(2)}_{exch-disp}$(CKS)	0.213
$E^{SAPT(DFT)}_{int}$	−6.006

[a] The cc-pVDZ basis set in the MC+BS form with 3*s*3*p*2*d*2*f* midbond was used. The unit of energy is kcal/mol. All calculations in double precision.

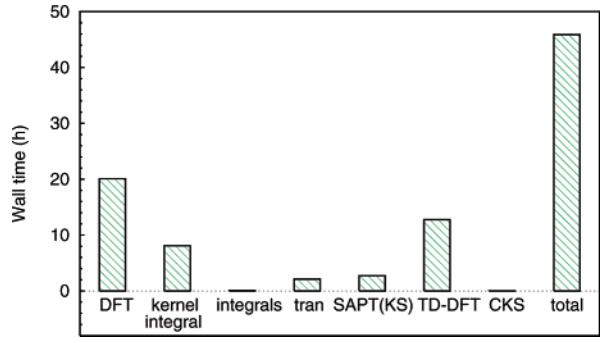


Figure 2. Wall times for the RDX dimer on 2.4 GHz Opteron. The cc-pVDZ basis set with 3*s*3*p*2*d*2*f* midbond was used corresponding to $o = 57$, $v = 366$, $N_{aux} = 1948$, $g = 742\,375$ grid points in the DFT and TD-DFT calculations. For the meaning of the symbols, see Figure 1.

in the CKS steps (TD-DFT matrix multiplications and the dispersion energy evaluation), as the scaling of these steps is reduced from the sixth to the fifth power of system size.

To demonstrate the capabilities of the present approach, we show the results for the RDX dimer in Table 4. Availability of accurate interaction potentials for systems such as RDX is crucial for first-principles predictions of properties of molecular crystals. Due to the size of this system ($o = 57$, $n = 423$), the use of the standard SAPT(DFT) algorithm turned out to be not possible with our current computer resources. Therefore, only the results with density-fitting are shown. Judging from the performance of this approach for other systems, also here the fitting error is most likely negligible and the presented interaction energy components are the most accurate ab initio results to date available for this system. It should be noted that SAPT(DFT) is currently the only practical method of accurately calculating the dispersion energy for systems of the size of RDX. Table 4 shows that, for the considered geometry, the latter component constitutes more than half of the total interaction energy and certainly cannot be neglected in any simulations of the crystal structure.

In Figure 2, the timings of our density-fitted calculations for the RDX dimer are presented. Although the monomer Kohn−Sham portion is still the most time-consuming, the $O(N^5)$ TD-DFT step is now seen to take a comparable amount of time. The cost of transformation, although nominally also scaling as $O(N^5)$, is relatively small due to a smaller prefactor

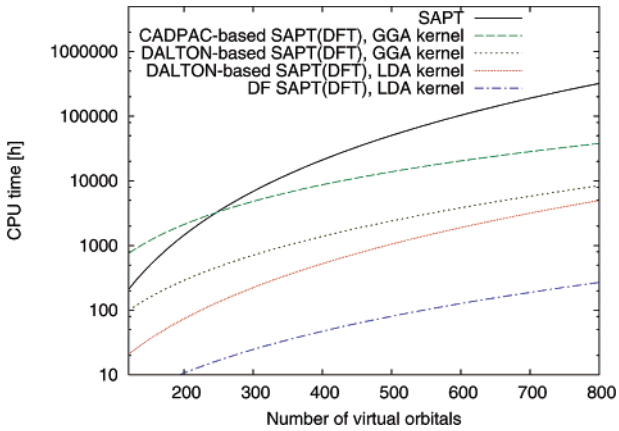


Figure 3. Estimates of the timings for the RDX dimer based on theoretical scaling and extrapolation of the data obtained for smaller systems. 'CADPAC-based SAPT(DFT)' refers to the implementation of ref 10 with no density fitting, 'DALTON-based SAPT(DFT)' is the version without density fitting, 'DF SAPT(DFT)'—density-fitting implementation of the present work and of ref 22, DALTON-based.

than in the case of the TD-DFT terms. Thus, for still larger systems, the latter step is going to dominate the whole calculation.

In Figure 3, we present a visualization of the overall numerical scalings of the SAPT-based methods. We assumed $o = 57$, as for the RDX dimer. The results of the graph were obtained by fitting timings of the various steps of the calculation for the dimethylnitroamine (DMNA) and benzene dimers. Although exact timings cannot be predicted in this way, the graph provides a qualitative comparison of the methods. The regular SAPT calculations are several orders of magnitude more time-consuming than density-fitted SAPT(DFT) ones and significantly more time-consuming than even the CADPAC-based SAPT(DFT) calculations, except for small basis sets where the latter suffer from an inefficient implementation of the integral 20 in CADPAC. Similar conclusions would hold for other correlated methods which include contributions from triple excitations, for example for the coupled cluster method with single, double, and noniterative triple excitations [CCSD(T)]. Our current, DALTON-based implementation is much faster than the former CADPAC-based one of ref 10 mainly due to the more efficient programing of integral 20. The use of the LDA kernel in integral 20 produces another important speedup with a minimal loss of accuracy, as shown in ref 10. Still, the largest relative speedup is due to the use of density fitting implemented in the present work and in ref 22. At the edge of the figure, i.e., for $v = 800$, SAPT calculations (or CCSD(T) calculations) would require 27 years of CPU time, whereas SAPT(DFT) calculations with density fitting take 11 days, a medium-size task if the work is distributed among a few dozen processors.

VI. Conclusions

We have presented a complete implementation of the SAPT(DFT) method based on density fitting of molecular integrals. The density-fitting approximation, applied at the stages of integral transformation, TD-DFT calculations, and in evalu-

ation of the CKS induction and dispersion energies, results in reductions of scaling and operation counts of these most time-consuming steps and hence offers a significant speedup over the standard formulation without density fitting. The overall time requirement of density-fitting SAPT(DFT) scales as $O(N^5)$ in contrast to the $O(N^6)$ scaling of the standard version. Moreover, the memory and IO-requirements of the algorithm are also greatly reduced. All these improvements enable high-accuracy studies of interactions in molecular complexes inaccessible to the standard, wave function-based ab initio methods. The interaction energy for an example of such a complex, the RDX dimer consisting of 42 atoms, 57 occupied orbitals, and using a basis set containing 423 functions, has been computed in this work. Although density-fitting introduces an error in the calculated interaction energies, we have shown that this error is very small, well below 1% for individual energy components. Our implementation is valid for both nonhybrid and hybrid density functionals. As shown in ref 22, in the latter case, an additional (besides density fitting of integrals) approximation has to be applied to bring the cost of obtaining the CKS propagators to $O(N^5)$ scaling. This approximation, consisting in truncation of an iterative scheme of solving the TD-DFT equations, does not significantly impair the overall accuracy.[22] Our new algorithms are flexible with respect to the type of basis sets, i.e., work with both dimer- and monomer-centered bases, including the monomer-centered 'plus' sets of ref 24, which are usually a good compromise between the size of the basis set and accuracy of the computed interaction energies. The implementation utilizes parts of the existing SAPT2002 suite of codes[23] and therefore can readily benefit from developments and updates made in this suite. In particular, the third-order code recently added to SAPT2002 can be used to extend SAPT(DFT). Work in this direction is in progress. The method can also be applied to three-body interactions.[55−57]

Acknowledgment. Funding for this work was provided by an ARO DEPSCOR grant and by the NSF grant CHE0239611.

Supporting Information Available: The RDX dimer geometry. This material is available free of charge via the Internet at http://pubs.acs.org.

References

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

(2) Jeziorski, B.; Szalewicz, K. In *Encyclopedia of Computational Chemistry*; von Ragué Schleyer, P., et al., Eds.; Wiley: Chichester, 1998; Vol. 2, pp 1376−1398.

(3) Jeziorski, B.; Szalewicz, K. In *Handbook of Molecular Physics and Quantum Chemistry*; Wilson, S., Ed.; Wiley: 2003; Vol. 3, Part 2, Chapter 9, pp 232−279.

(4) Williams, H. L.; Chabalowski, C. F. *J. Phys. Chem. A* **2001**, *105*, 646−659.

(5) Misquitta, A. J.; Szalewicz, K. *Chem. Phys. Lett.* **2002**, *357*, 301−306.

(6) Hesselmann, A.; Jansen, G. *Chem. Phys. Lett.* **2002**, *357*, 464−470.

(7) Hesselmann, A.; Jansen, G. *Chem. Phys. Lett.* **2002**, *362*, 319−325.

(8) Misquitta, A. J.; Jeziorski, B.; Szalewicz, K. *Phys. Rev. Lett.* **2003**, *91*, 033201.

(9) Hesselmann, A.; Jansen, G. *Chem. Phys. Lett.* **2003**, *367*, 778−784.

(10) Misquitta, A. J.; Podeszwa, R.; Jeziorski, B.; Szalewicz, K. *J. Chem. Phys.* **2005**, *123*, 214103.

(11) Misquitta, A. J.; Szalewicz, K. *J. Chem. Phys.* **2005**, *122*, 214109.

(12) Szalewicz, K.; Podeszwa, R.; Misquitta, A. J.; Jeziorski, B. In *Lecture Series on Computer and Computational Science. ICCMSE 2004*; Simos, T., Maroulis, G., Eds.; VSP: Utrecht, 2004; Vol. 1, pp 1033−1036.

(13) Podeszwa, R.; Szalewicz, K. *Chem. Phys. Lett.* **2005**, *412*, 488−493.

(14) Baerends, E. J.; Ellis, D. E.; Ros, P. *Chem. Phys.* **1973**, *2*, 41−51.

(15) Sambe, H.; Felton, R. *J. Chem. Phys.* **1975**, *62*, 1122−1126.

(16) Dunlap, B. I.; Connolly, J. W. D.; Sabin, J. R. *J. Chem. Phys.* **1979**, *71*, 4993−4999.

(17) Jamorski, C.; Casida, M. E.; Salahub, D. R. *J. Chem. Phys.* **1995**, *104*, 5134−5147.

(18) Dunlap, B. I. *Phys. Chem. Chem. Phys.* **2000**, *2*, 2113−2116.

(19) Werner, H.-J.; Manby, F. R.; Knowles, P. J. *J. Chem. Phys.* **2003**, *118*, 8149−8160.

(20) Hesselmann, A.; Jansen, G.; Schütz, M. *J. Chem. Phys.* **2005**, *122*, 014103.

(21) Della Sala, F.; Görling, A. *J. Chem. Phys.* **2001**, *115*, 5718−5732.

(22) Bukowski, R.; Podeszwa, R.; Szalewicz, K. *Chem. Phys. Lett.* **2005**, *414*, 111−116.

(23) *SAPT2002: An Ab Initio Program for Many-Body Symmetry-Adapted Perturbation Theory Calculations of Intermolecular Interaction Energies*; by Bukowski, R.; Cencek, W.; Jankowski, P.; Jeziorski, B.; Jeziorska, M.; Kucharski, S. A.; Lotrich, V. F.; Misquitta, A. J.; Moszynski, R.; Patkowski, K.; Rybak, S.; Szalewicz, K.; Williams, H. L.; Wormer, P. E. S. University of Delaware and University of Warsaw (http://www.physics.udel.edu/∼szalewic/SAPT/SAPT.html).

(24) Williams, H. L.; Mas, E. M.; Szalewicz, K.; Jeziorski, B. *J. Chem. Phys.* **1995**, *103*, 7374−7391.

(25) Longuet-Higgins, H. C. *Discuss. Faraday Soc.* **1965**, *40*, 7−18.

(26) Zaremba, E.; Kohn, W. *Phys. Rev. B* **1976**, *13*, 2270−2285.

(27) McWeeny, R. *Croat. Chem. Acta* **1984**, *57*, 865−878.

(28) Angyan, J. G.; Jansen, G.; Loos, M.; Hattig, C.; Hess, B. A. *Chem. Phys. Lett.* **1994**, *219*, 267−273.

(29) DALTON, a molecular electronic structure program, Release 2.0 2005, see http://www.kjemi.uio.no/software/dalton/dalton.html.

(30) CADPAC: The Cambridge Analytic Derivatives Package Issue 6, Cambridge, 1995. A suite of quantum chemistry programs developed by R. D. Amos with contributions from I. L. Alberts et al.

(31) Tozer, D. J.; Handy, N. C. *J. Chem. Phys.* **1998**, *109*, 10180−10189.

(32) Tozer, D. J.; Amos, R. D.; Handy, N. C.; Roos, B. O.; Serrano-Andrés, L. *Mol. Phys.* **1999**, *97*, 859−868.

(33) Lee, A. M.; Colwell, S. M. *J. Chem. Phys.* **1994**, *101*, 9704−9709.

(34) Ioannou, A. G.; Colwell, S. M.; Amos, R. D. *Chem. Phys. Lett.* **1997**, *278*, 278−284.

(35) Colwell, S. M.; Handy, N. C.; Lee, A. M. *Phys. Rev. A* **1996**, *53*, 1316−1322.

(36) Colwell, S. M.; Murray, C. W.; Handy, N. C.; Amos, R. D. *Chem. Phys. Lett.* **1993**, *210*, 261−268.

(37) Schmidt, M. W.; Baldridge, K. K.; Boatz, J. A.; Elbert, S. T.; Gordon, M. S.; Jensen, J. H.; Koseki, S.; Matsunaga, K. A. N.; Su, S. J.; Windus, T. L.; Dupuis, M.; Montgomery, J. A. *J. Comput. Chem.* **1993**, *14*, 1347−1363.

(38) Casida, M. E. In *Recent Advances in Density-Functional Theory Part I, Time-Dependent Density Functional Response Theory for Molecules*; Chong, D. P., Ed.; World Scientific: Singapore, 1995; pp 155−192.

(39) Amos, R. D.; Handy, N. C.; Knowles, P. J.; Rice, J. E.; Stone, A. J. *J. Phys. Chem.* **1985**, *89*, 2186−2192.

(40) van Gisbergen, S. J. A.; Snijders, J. G.; Baerends, E. J. *Comput. Phys. Comm.* **1999**, *118*, 119−138.

(41) Jeziorski, B.; Moszynski, R.; Ratkiewicz, A.; Rybak, S.; Szalewicz, K.; Williams, H. L. In *Methods and Techniques in Computational Chemistry: METECC-94*; Clementi, E., Ed.; STEF: Cagliari, 1993; Vol. B, pp 79−129.

(42) Patkowski, K.; Jeziorski, B.; Szalewicz, K. *J. Chem. Phys.*, submitted.

(43) Whaley, R. C.; Petitet, A.; Dongarra, J. J. *Parallel Comput.* **2001**, *27*, 3−35.

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

(45) Adamo, C.; Barone, V. *J. Chem. Phys.* **1999**, *110*, 6158−6170.

(46) Fermi, E.; Amaldi, E. *Mem. Accad. Italia* **1934**, *6*, 119−149.

(47) Lias, S. G. Ionization Energy Evaluation, in NIST Chemistry WebBook, NIST Standard Reference Database Number 69 (http://webbook.nist.gov).

(48) Kendall, R. A.; Dunning, T. H.; Harrison, R. J. *J. Chem. Phys.* **1992**, *96*, 6796−6806.

(49) Weigend, F.; Köhn, A.; Hättig, C. *J. Chem. Phys.* **2002**, *116*, 3175−3183.

(50) Tsuzuki, S.; Honda, K.; Mikami, M.; Tanabe, K. *J. Am. Chem. Soc.* **2002**, *124*, 104−112.

(51) Weigend, F. *Phys. Chem. Chem. Phys.* **2002**, *4*, 4285−4291.

(52) Cisneros, G. A.; Piquemal, J.-P.; Darden, T. A. *J. Chem. Phys.* **2005**, *123*, 044109.

(53) Choi, C. S.; Prince, E. *Acta Crystallogr.* **1972**, *B 28*, 2857−2862.

(54) See the Supporting Information.

(55) Lotrich, V. F.; Szalewicz, K. *J. Chem. Phys.* **1997**, *106*, 9668−9702.

(56) Lotrich, V. F.; Szalewicz, K. *Phys. Rev. Lett.* **1997**, *79*, 1301−1304.

(57) Lotrich, V. F.; Szalewicz, K. *J. Chem. Phys.* **2000**, *112*, 112−121.

CT050304H

JCTC Journal of Chemical Theory and Computation

NO-MNDO: Reintroduction of the Overlap Matrix into MNDO

Kurt W. Sattelmeyer, Ivan Tubert-Brohman, and William L. Jorgensen*

Department of Chemistry, Yale University, 225 Prospect Street, New Haven, Connecticut 06520

Received July 18, 2005

Abstract: The effect of reintroducing the overlap matrix into the secular equations for an NDDO (neglect of diatomic differential overlap)-based semiempirical molecular orbital method has been investigated. The modification is expected to improve the description of interactions between electron pairs. The idea has been tested by implementation and evaluation of a nonorthogonal version of the MNDO method (NO-MNDO) with parametrization for hydrogen, carbon, nitrogen, and oxygen. Overall, the accuracy of NO-MNDO for heats of formation is nearly identical to that for the more highly parametrized AM1 method. The mean absolute error (MAE) for heats of formation of a comprehensive set of 622 neutral, closed-shell molecules is reduced from 8.4 kcal/mol with MNDO to 6.8 kcal/mol with NO-MNDO. In addition, the performance for conformational equilibria and torsional barriers is significantly improved with NO-MNDO, presumably owing to the improved description of closed-shell interactions. For molecular geometries, the usual training and test sets have been expanded through use of MP2/6-31G(d) results for consistent comparisons. The performance of NO-MNDO for bond lengths, bond angles, and dihedral angles remains good with MAEs of 0.017 Å, 2.5°, and 4.5°. Additionally, NO-MNDO corrects severe errors by MNDO for R$^\bullet$ + H−R′ hydrogen-atom transfers, while testing for activation barriers for nine pericyclic reactions reveals only modest improvement.

1. Introduction

The speedup afforded by the Neglect of Diatomic Differential Overlap (NDDO)[1−3] approximation has made the semiempirical molecular orbital (SMO) methods based on it, including MNDO,[4,5] AM1,[6] PM3,[7,8] and MNDO/d,[9] valuable tools when a more rigorous approach is precluded by either the size of the system or the number of computations required. While mean absolute errors (MAEs) using these schemes do not reach chemical accuracy (ca. 1.0 kcal/mol, as can be approached using the best available *N*[7] ab initio methods, such as CCSD(T)[10,11]), the structures and energetics from SMO methods are often acceptable for many applications. For example, MNDO, AM1, and PM3 give MAEs for heats of formation of 6.8, 5.1, and 4.1 kcal/mol for the 56 molecules in the combined G2-1 and G2-2 sets,[12,13] which contain only H, C, N, and O atoms. Nevertheless, SMO methods suffer from a number of problems. Common errors include prediction of straight chain hydrocarbons to be more stable than branched isomers, underestimation of rotational barriers, overestimation of activation energies for pericyclic reactions, and significant energetic errors for molecules containing adjacent heteroatoms or small rings.[14,15]

Recent attempts to improve upon these semiempirical methods have largely centered upon corrections to the core repulsion formula (CRF). Specifically, the Pairwise Distance Directed Gaussian (PDDG)[16] extension in eq 1 to the MNDO and PM3 CRFs and subsequent reparametrizations

$$\mathrm{PDDG}(A,B) = \mathrm{CRF}_{\mathrm{MNDO}} + \frac{1}{n_A + n_B}\left[\sum_{i=1}^{2}\sum_{j=1}^{2}(n_A P_{Ai} + n_B P_{Bj})\mathrm{e}^{-10\text{Å}^{-2}(R_{AB}-D_{Ai}-D_{Bj})^2}\right] \quad (1)$$

yield large improvements for heats of formation and isomerization energies without significantly degrading other proper-

* Corresponding author e-mail: william.jorgensen@yale.edu.

10.1021/ct050174c CCC: $33.50
Published on Web 02/22/2006

ties. The PDDG correction to MNDO lowers the MAEs for a comprehensive set of 622 neutral, closed-shell molecules containing only H, C, N, and O from 8.4 to 5.2 kcal/mol. It is not suprising that this result is somewhat worse than the 4.4 kcal/mol MAE of PM3 (the CRF for PM3 is represented in eq 2), as the number of parameters per atom for the added Gaussians has been reduced

$$\mathrm{PM3}(A,B) = \mathrm{CRF}_{\mathrm{MNDO}} + \frac{Z_A Z_B}{R_{AB}}\left(\sum_{i=1}^{2} a_{Ai} e^{-b_{Ai}(R_{AB}-c_{Ai})^2} + \sum_{i=1}^{2} a_{Bi} e^{-b_{Bi}(R_{AB}-c_{Bi})^2}\right) \quad (2)$$

from six to four, with the discrepancy arising from the *b* parameters in eq 2 being taken as constants in PDDG/MNDO. It is also not surprising that the combined PDDG/PM3 method gives the lowest MAE, 3.2 kcal/mol. Additionally, Thiel et al. have pursued a different approach in their OM1 and OM2 methods[17] through modifications to the NDDO version of the Roothann−Hall equations.[18,19] In their work, the lack of Pauli repulsion has been addressed by the addition of three-center terms into the two-center core Hamiltonian, yielding a better description of barriers to internal rotation.

While the these methods have been largely successful in addressing certain problems with specific implementations of the NDDO formalism, ad hoc modifications have thus far not led to a general purpose method capable of correcting all of the problems simultaneously. With this in mind, we have been exploring simple modifications to yield a general, improved SMO method that can be easily implemented in existing SMO codes and that is particularly appropriate for QM/MM calculations, while still retaining the favorable N^3 scaling. In this paper, one such modification, namely, an alternative treatment of the Pauli repulsion issue, is considered by reintroduction of the overlap matrix into the secular equations.

2. Nonorthogonalized MNDO (NO-MNDO) Formalism

Much of the motivation behind Thiel's OM1 and OM2 is the improper, even splitting of mixing orbitals that arises from neglecting the overlap integrals in the secular equations for ZDO (zero differential overlap) theories and by extension for the NDDO-based methods. Instead, the lower-lying orbital should fall by less than the higher-lying orbital increases in energy. Thus, when both orbitals are doubly occupied, there is an intrinsic repulsion, the "Pauli repulsion," which increases as the overlap of the orbitals increases. This has nothing to do with two-electron integrals and simply results from deriving secular equations for one- or many-electron systems. Neglect of Pauli repulsion can be expected to contribute to the underestimation of rotational barriers and problems with the treatment of adjacent heteroatoms with lone pairs and electronic excitation energies, which are common with SMO methods. Although Thiel's OM methods represent a reasonable palliation, the orthogonalization issue has been addressed here more simply by reintroducing the overlap matrix (**S**) into the secular equations. Therefore, we have returned to solving **FC=SCE**. This modification does not increase the scaling of the method or the number of parameters versus MNDO. The **S** matrix is always available in SMO calculations as it is used in the computation of the one-electron, two-center resonance integrals, $\beta_{\mu\nu}$. In calculating heats of formation, the PDDG approach has also been used here, i.e., the electronic energy of an atom (*eisol*) is treated as an optimizable parameter and not as one derived from calculations on the atom. This parameter is simply set to minimize the MAEs.

Table 1. Molecular Properties Used in Construction of the Error Function for Parameter Optimization

data type	*N*	weighting factor
heat of formation	355	1 mol/kcal
bond length	153	100 $Å^{-1}$
bond angle	93	2/3 deg^{-1}
dihedral angle	15	1/3 deg^{-1}
ionization potential	66	10 eV^{-1}
dipole moment	42	20 D^{-1}

It can be speculated that the motivation to remove the overlap terms from the secular equations for the seminal ZDO-based SMO method, CNDO,[20] by Pople et al. in 1965 was somewhat influenced by practicality since the requisite second matrix diagonalization in solving the nonorthogonal eigenvalue problem essentially doubles the required computer time. Though overlap distributions $\varphi_\mu\varphi_\nu$ were neglected in two-electron integrals, they were, of course, never neglected for one-electron, two-center resonance integrals, and their neglect in the secular equations was arbitrary from a theoretical standpoint and physically incorrect. Furthermore, it should be noted that simple Hückel theories also neglect the overlap integrals in the secular equations, while they are included in Hoffmann's extended Hückel method (EHT) from 1963.[21] The practical difference here is that EHT calculations are noniterative, so the two matrix diagonalizations are only performed once, while SMO calculations require a normal SCF cycle. The inclusion of Pauli repulsion in EHT has been known for many years to be essential to its qualitative success in describing orbital interactions, aromaticity, and rotational barriers.[22]

3. Optimization of the Parameters

The impact of the reintroduction of the overlap integrals into the secular equations has been tested by (1) making the modification to the MNDO method to yield nonorthogonal MNDO (NO-MNDO), (2) parametrizing the method for molecules containing C, H, N, and O atoms, and (3) comparing the results with those from MNDO and other SMO methods. The required computer time for a NO-MNDO calculation is ca. twice that for an MNDO calculation owing to the second matrix diagonalization, as expected. It should be noted that NO-MNDO is not a method that we intend to utilize further; it was simply pursued to gauge the importance of the orthogonality issue for future SMO development.

All original MNDO parameters (U_{ss}, β_s, ζ_s, and α for hydrogen and U_{ss}, U_{pp}, β_s, β_p, ζ_s, ζ_p, and α for carbon, nitrogen, and oxygen) were reoptimized. As previously explained[16,23] the optimization process consists of three

Table 2. MNDO and NO-MNDO Parameters[a]

	MNDO[b]				NO-MNDO			
	H	C	N	O	H	C	N	O
U_{ss}	−11.906276	−52.279745	−71.932122	−99.64309	−10.880363	−50.189763	−69.782951	−96.705658
U_{pp}		−39.205558	−57.172319	−77.797472		−39.547267	−56.981889	−76.391762
β_s	−6.989064	−18.985044	−20.495758	−32.688082	−9.364858	−16.208034	−24.905520	−35.477596
β_p		−7.934122	−20.495758	−32.688082		−10.637421	−21.291958	−28.881783
ζ_s	1.331967	1.787537	2.255614	2.699905	1.061597	1.925428	2.351138	2.455548
ζ_p		1.787537	2.255614	2.699905		1.727933	1.951819	2.537964
α	2.544134	2.546380	2.861342	3.160604	2.687705	2.484460	2.658599	2.946645
eisol	−11.906276	−120.500606	−202.566201	−317.868506	−13.160122	−119.594403	−202.243601	−304.341294
DD		0.807466	0.639904	0.534602		0.784403	0.656525	0.577707
QQ		0.685158	0.542976	0.453625		0.708792	0.627489	0.482570
ρ_0[c]	1.058920	1.112429	1.001103	0.882296	1.058920	1.112428	1.001103	0.882296
ρ_1[c]		0.813078	0.637459	0.521237		0.800239	0.646434	0.543492
ρ_2[c]		0.747842	0.615275	0.526541		0.765271	0.679156	0.549476

[a] Units are as follows: (eV) U_{ss}, U_{pp}, β_s, β_p, eisol; (au) ζ_s, ζ_p; (Bohr) *DD, QQ*, ρ_0, ρ_1, ρ_2; (Å) α. [b] References 4 and 5. [c] For use in MOPAC 6, $\rho_0 = 0.5/AM$, $\rho_1 = 0.5/AD$, $\rho_2 = 0.5/AQ$.

stages. First, random displacements of the parameters are generated, and simulated annealing is used to minimize an error function constructed from the properties listed in Table 1 and from the gradients for bond lengths, bond angles, and dihedral angles on the nonoptimized structures of a training set of 126 molecules. All reference geometry parameters have now been taken from MP2/6-31G* calculations, which allowed a large expansion in the size of the training set. Additionally, heats of formation for a total of 355 molecules are present in this training set, including heats of formation of several transition state structures. Subsequently, promising parameter sets (as determined by their small error functions) are optimized with full geometry optimization for all molecules, neglecting the (zero) contribution of gradients to the error function. Final values of E_{el}^{A} or "*eisol*", the electronic energy for each element *A*, are determined by minimizing the MAE in heats of formation for 473 molecules (464 neutral, closed-shell molecules and nine transition states of pericyclic reactions). Equation 3 provides the relationship between a molecule's ΔH_{f}, electronic energy, E_{mol}, and E_{el}^{A}.

$$\Delta H_{\mathrm{f}} = E_{\mathrm{mol}} + \sum_{A} (\Delta H_{\mathrm{f}}^{A} - E_{\mathrm{el}}^{A}) \qquad (3)$$

The final testing for heats of formation was carried out on the full set of 622 molecules that was used previously.[16] All SMO calculations have been executed with a local version of MOPAC 6.[24] Detailed results for all molecules and transition structures are presented in the Supporting Information.

4. Results and Discussion

Energetics. Table 2 lists the newly optimized parameters for NO-MNDO as well as those for MNDO, and Table 3 shows the performance of NO-MNDO and the other semiempirical methods for heats of formation. As expected, due to the global search method employed and the change in methodology, a significantly different parameter emerged. The most striking change is in the β_s value for hydrogen, which is more than 30% lower in NO-MNDO. Overall, the 6.8 kcal/mol MAE for NO-MNDO represents a 1.6 kcal/mol improvement over MNDO. NO-MNDO shows greater than 2 kcal/mol improvements for CH and CHO containing compounds, but it currently does less well than MNDO with CHN containing compounds. It is possible that further parameter search would correct this anomaly. However, in comparison to MNDO, NO-MNDO benefits from the optimization of the *eisol* values as well as the inclusion of the overlap matrix. Specifically, we previously found that optimization of the *eisol* values in conjunction with a complete reoptimization of the other MNDO parameters yields an MNDO version with an MAE of 7.3 kcal/mol for the 622 molecules.[16] Thus, 0.5 kcal/mol of the remaining error is removed in proceeding to the current NO-MNDO. This is encouraging, particularly if a similar gain could be made starting from PM3, as the new method retains the same scaling properties without introducing any new optimizable parameters. It is also notable that nearly identical MAEs are obtained with NO-MNDO and AM1; however, two to four Gaussian functions per element are added to the core repulsion formula for AM1 along with 36 additional parameters for coverage of C, H, N, and O. It is apparent that much of the success of AM1 over MNDO comes from the reparametrization including independent optimization of the orbital exponents ζ_s and ζ_p rather than from the addition of the Gaussians to the CRF. The optimization of MNDO was clearly constrained by the modest computer resources available in the mid-1970s.[4,5]

Table 3. Mean Absolute Errors for Heats of Formation of Neutral, Closed-Shell Molecules (kcal/mol)

		standard NDDO			PDDG		nonorthog
molecules	*N*	MNDO	AM1	PM3	MNDO	PM3	NO-MNDO
all	622	8.4	6.7	4.4	5.2	3.2	6.8
HC	254	8.0	5.6	3.6	5.1	2.6	5.8
HCN	89	6.3	7.3	4.7	5.7	4.2	10.5
HCO	238	8.7	7.2	4.6	5.0	3.2	6.2
HCNO	41	13.4	9.5	7.0	4.9	4.5	9.3

A related gauge of success is the relative abundance of outliers, i.e., molecules for which the computed heat of

Table 4. Number of Molecules (out of 622) Computed to Have Heats of Formation Differing from Experimental Values by More than 15 and 30 Kcal/Mol

	standard NDDO			PDDG		nonorthog
	MNDO	AM1	PM3	MNDO	PM3	NO-MNDO
> 15 kcal/mol	97	49	16	20	7	55
> 30 kcal/mol	18	7	2	4	1	4

formation differs from the experimental reference value by more than some large value. The numbers of these are listed in Table 4 using the arbitrary cutoffs of 15 and 30 kcal/mol. Again, NO-MNDO demonstrates a significant improvement over MNDO and is similar in performance to AM1. As detailed in Table 5, out of the 622 minimum-energy structures considered in this work, NO-MNDO is not able to reproduce the experimental heat of formation to within 30 kcal/mol for diazirine, carbon suboxide, isophthalamide, and N_2. Stewart has recently suggested that there is an error in the experimentally reported heat of formation of isophthalamide,[25] which is also greatly overpredicted by AM1 and PM3. Using his value of −70.3 kcal/mol reduces the errors for isophthalamide by 21.1 kcal/mol. The other poorest performing cases for NO-MNDO are all small and have unique bonding characteristics. Cubane, which suffers from the additive errors of multiple four-membered rings, populates Table 5 for all SMO methods except AM1 and NO-MNDO. In general, the greatest problems for NO-MNDO occur with acetylenes, nitrogen-containing aromatic heterocycles, and compounds containing nitrogen−nitrogen multiple bonds.

A well-known problem for MNDO and AM1 is that they erroneously find branched isomers to be less stable than straight-chain ones. MNDO's most severe branching problems (tri-*tert*-butyl methane, 2,2,3,3-tetramethylpentane, 2,3,3,4-tetramethylpentane, etc.) are largely corrected in NO-MNDO by the use of a larger α in the CRF for hydrogen. However, there is a fine balance here with the limited number of parameters available, as choosing too large an exponent results in a marked contraction of H−C bond lengths. Less extreme cases, such as the pentane/neopentane enthalpy difference of −5.0 kcal/mol, also show some improvement with NO-MNDO. MNDO and AM1 predict +9.8 and +5.2 kcal/mol, while NO-MNDO yields +3.5 kcal/mol. PM3 does better at −1.3 kcal/mol, while the expanded core repulsion formula with PDDG/PM3 adequately solves the problem (−7.2). For butane vs isobutane, NO-MNDO gives +0.8 kcal/mol, while the experimental, MNDO, AM1, PM3, and PDDG/PM3, numbers are −2.0, +2.9, +1.7, −0.4, and −2.5 kcal/mol, respectively.

Another area of potential improvement for SMO methods is in the description of the transition-state energetics of prototypical pericyclic reactions, especially in comparison to their overall treatment of hydrocarbons. The consistent trend is an overestimation of the reaction barriers. As shown in Table 6, AM1 is best able to reproduce the experimental barriers for the nine representative reactions, which have been thoroughly studied by Guner et al.,[26,27] giving an average error of 6.8 kcal/mol. In comparison, HF/6-31G*, MP2/6-31G*, KMLYP/6-31G*, and B3LYP/6-31G* have MAEs of 18.7, 4.6, 3.2, and 1.7 kcal/mol, respectively. Due to the current deficiencies, the activation barriers for these nine prototypical pericyclic reactions were explicitly included in the parametrization of NO-MNDO. As demonstrated in Table 6, this led to only modest improvement over MNDO, even though the contribution of these errors to the overall error

Table 5. Problematic Heats of Formation and Their Differences from Experimental Values

standard NDDO			PDDG		nonorthog
MNDO	AM1	PM3	MNDO	PM3	NO-MNDO
ozone (+94.6)	isophthalamide (+38.2)	isophthalamide (+35.8)	ozone (−51.4)	cubane (−39.1)	diazirine (+60.4)
tri-*tert*-butylmethane (+88.8)	di-*tert*-butyl peroxide (+34.3)	cubane (−34.9)	isobutylamine (−45.4)		carbon suboxide (−50.6)
cubane (−49.6)	bicyclo[1.1.1] pentane (+33.4)		diadamantanone (+41.6)		isophthalamide (+41.8)
3,3,4,4-tetramethyl-2-pentanone (+45.7)	5-methylisoxazole (+31.8)		cubane (−39.2)		nitrogen (+39.2)
di-*tert*-butyl peroxide (+45.1)	3,5-dimethylisoxazole (+31.7)				

Table 6. Activation Enthalpies for Selected Pericyclic Reactions (kcal/mol)

		standard NDDO			PDDG		nonorthog
reaction	expt[a]	MNDO	AM1	PM3	MNDO	PM3	NO-MNDO
cyclobutene opening	31.9	49.8	35.3	40.6	44.8	41.3	51.3
1,3,5-hexatriene closure	30.2	40.0	31.0	31.2	38.7	36.6	46.3
o-xylylene closure	28.1	40.3	38.8	38.9	43.6	41.2	46.4
1,3-pentadiene [1,5]-H shift	36.8	57.4	39.8	36.4	46.1	32.4	39.1
cyclopentadiene [1,5]-H shift	23.7	48.9	39.5	37.7	47.5	30.7	42.1
1,5-hexadiene Cope	34.5	40.9	37.6	41.8	42.8	45.9	53.8
ethylene + 1,3-butadiene DA	25.0	45.3	23.8	27.0	41.1	30.0	44.0
ethylene + cyclopentadiene DA	23.7	50.5	28.5	32.1	42.1	33.0	51.0
cyclopentadiene dimerization	15.9	50.0	34.2	37.4	44.3	38.3	26.6
MAE		19.3	6.8	8.2	15.7	9.8	16.7

[a] References 26 and 27.

Table 7. Barrier Heights for Hydrogen Transfer Reactions (kcal/mol)

		standard NDDO[a]			PDDG[a]		
	consensus[a]	MNDO	AM1	PM3	MNDO	PM3	nonorthog NO-MNDO
$CH_3^\bullet + CH_4$	17.53	28.59	13.49	10.14	20.36	6.51	9.55
$CH_3^\bullet + C_2H_6$	15.36	79.59	12.01	7.29	18.71	3.45	6.92
$C_2H_5^\bullet + CH_4$	18.99	85.89	17.81	15.60	25.17	12.07	16.45
$C_2H_5^\bullet + C_2H_6$	16.69	32.15	16.02	11.99	23.41	8.29	13.08
$C_3H_7^\bullet + C_3H_8$	16.04	32.45	15.56	12.47	22.80	7.33	16.89
MAE		34.8	1.9	5.4	5.2	9.4	4.7

[a] Reference 28.

Table 8. Conformational and Isomerization Energies (kcal/mol) for Prototypical Organic Molecules

			standard NDDO			PDDG		
molecule		ref ΔE	MNDO	AM1	PM3	MNDO	PM3	nonorthog NO-MNDO
butane (trans)	skew	3.6[a]	1.4	1.5	1.6	2.2	1.6	1.0
	gauche	0.7	0.6	0.7	0.6	1.3	0.3	−0.4
	cis	5.7	3.2	3.3	4.0	5.5	3.9	4.4
ethane (staggered)	eclipsed	2.8[b]	1.0	1.2	1.4	2.2	1.1	1.4
methylcyclohexane (equatorial)	axial	1.8[c]	6.6	1.4	1.1	4.4	0.9	0.3
propene (eclipsed)	bisected	2.0[d]	0.2	0.6	0.7	0.1	0.7	1.0
2-butene (trans)	cis	1.0[e]	0.8	1.1	0.2	0.8	1.5	0.9
1,3-butadiene (trans)	skew	2.49[e]	0.3	0.8	0.7	1.4	0.7	1.7
1-butene (skew)	cis	0.53[e]	1.3	0.7	0.9	2.3	0.7	0.8
propanal (cis)	skew	0.95[e]	−0.3	−0.6	−0.7	−1.5	−1.1	0.6
N-methylacetamide (Z)	E	2.3[e]	1.0	1.6	0.4	1.3	1.9	2.2
acrolein (trans)	cis	2.0[e]	−0.4	0.2	0.4	0.7	0.8	0.3
methyl formate (Z)	E	4.75[e]	2.9	5.6	1.9	0.2	1.8	6.7
MAE		0	1.8	1.1	1.4	1.5	1.3	1.1

[a] References 29 and 30. [b] References 31−34. [c] References 35 and 36. [d] Reference 37. [e] Reference 38.

function is not small. This suggests that the standard MNDO formalism does not allow enough flexibility to describe these activation barriers accurately. Indeed, inspection of the core repulsion formula of AM1 for carbon reveals two attractive Gaussians centered at 2.05 and 2.65 Å, thereby allowing AM1 to perform better due to the more favorable C−C interactions in this range. This result is somewhat spurious, though, as AM1 benefits from erroneous overestimation of the heats of formation of the reactants in these reactions; it is not noticeably better for the overall heats of reaction, giving an MAE of 7.1 kcal/mol for the six reactions with nonzero enthalpy changes compared to 8.6, 4.8, 3.9, 3.3, and 8.0 kcal/mol for MNDO, PM3, PDDG/MNDO, PDDG/PM3, and NO-MNDO, respectively.

We also examined the barrier heights of several hydrogen transfer reactions. The results are listed in Table 7 and accompany the recent, best estimates from Dybala-Defratyka et al.[28] In these cases, NO-MNDO does very respectably with an MAE of only 4.7 kcal/mol. The largest error is for the $CH_3^\bullet + C_2H_6$ reaction, where the barrier is underestimated by 8.4 kcal/mol. Furthermore, NO-MNDO is seen to correct the serious problems in the MNDO results, indicating that the standard MNDO formalism augmented with the Pauli repulsions is adequate here.

As a final energetic issue, Table 8 compares conformational and isomerization energies for prototypical molecules from the SMO methods with experimental and high-level, computed values from the literature.[29−38] NO-MNDO and AM1 perform the best among the SMO methods. Although the gauche structure of butane is predicted to be 0.4 kcal/mol more stable than anti using NO-MNDO, the overall improvement versus MNDO is apparent. While the anti to cis energy difference for butane is underestimated using MNDO by 2.5 kcal/mol and the energy difference between equatorial and axial methylcyclohexane is overestimated by 4.8 kcal/mol, the errors with NO-MNDO are just 1.3 and 1.5 kcal/mol, respectively. NO-MNDO is also the only method to give the cis structure of propanal as the minimum, as the other SMO methods have the skew structure lower by 0.3−1.5 kcal/mol. Thus, it seems likely that the addition of the Pauli repulsions in NO-MNDO has helped in this area.

Table 9. Mean Absolute Errors in Bond Lengths (Å)

		standard NDDO			PDDG		
	N	MNDO	AM1	PM3	MNDO	PM3	nonorthog NO-MNDO
training set	153	0.013	0.016	0.012	0.014	0.011	0.016
test set	65	0.018	0.019	0.013	0.020	0.017	0.020
all	218	0.015	0.017	0.012	0.016	0.013	0.017

Structure. Despite the improvements of NO-MNDO over MNDO with respect to energetics, it is important that the results for molecular geometries remain reasonable. In fact, as summarized in Tables 9−11, the overall quality of geometrical results is similar for all of the SMO methods. As previously mentioned, this work uses results of geometry optimizations at the MP2/6-31G* level for the reference values. This allowed expansion of the training set for bond lengths, bond angles, and dihedral angles, and comparisons can now be made in a more consistent manner, e.g., using

Table 10. Mean Absolute Errors in Bond Angles (deg)

		standard NDDO			PDDG		nonorthog
	N	MNDO	AM1	PM3	MNDO	PM3	NO-MNDO
training set	93	1.8	1.4	1.7	2.5	1.8	2.5
test set	33	2.0	1.7	1.6	2.3	1.9	2.4
all	126	1.9	1.5	1.7	2.4	1.9	2.5

Table 11. Mean Absolute Errors in Dihedral Angles (deg)

		standard NDDO			PDDG		nonorthog
	N	MNDO	AM1	PM3	MNDO	PM3	NO-MNDO
training set	15	4.6	2.3	2.8	2.8	3.5	4.5
test set	19	3.1	3.2	3.5	4.9	3.9	4.4
all	34	3.8	2.8	3.2	4.0	3.7	4.5

r_e values for bond lengths and not a collection of r_0, r_s, etc., depending on availability.

The greatest bond-length errors for NO-MNDO are for triple bonds. Those present in acetylene, propyne, isocyanomethane, 2-butyne, 1,3-butadiyne, vinylacetylene, and cyanogen are each underestimated by between 0.026 and 0.048 Å. In contrast, with an average error of 0.014 Å, the lengths of carbon−carbon single bonds are generally accurate and do not show systematic discrepancies. The largest errors here include an overestimate of the carbon−carbon single bonds of azirane by 0.046 Å and underestimates of those for propyne and 2-butyne by 0.041 and 0.040 Å. Other substantial deviations include underestimations of the bond length in molecular hydrogen by 0.099 Å and the nitrogen−nitrogen bond in hydrazine by 0.067 Å.

While NO-MNDO yields an MAE for bond angles slightly higher than the other SMO methods, it is seen to have its significant errors in many of the same situations as MNDO and PDDG/MNDO, specifically, the overestimation of angles where the central atom is oxygen. For example, the MNDO, PDDG/MNDO, NO-MNDO, and MP2/6-31G* COH bond angles for formic and acetic acid are 116.2°, 123.5°, 118.5°, and 106.1° and 115.6°, 122.9°, 117.8°, and 105.4°, respectively, while those for the COC angle of methyl formate are 125.7°, 127.9°, 121.2°, and 113.9°. Table 11 shows that the dihedral angle results from NO-MNDO are comparable to those of the other methods. AM1 appears to be the best performer in this area, though neither the data set nor the margin is large. The greatest sources of errors are also consistent across the methods and correspond to cases with relatively flat torsional energy surfaces.

Ionization Potentials and Dipole Moments. Ionization potentials from Koopman's theorem and dipole moments are also traditionally examined in papers reporting SMO methods. Though these properties were not emphasized in this study, the results with NO-MNDO compare reasonably well with those from the alternative SMO methods. The increase of the one-electron energy, U_{ss}, for hydrogen in Table 2 causes ionization potentials to generally be underestimated by ca. 1−2 eV for hydrocarbons with NO-MNDO; the largest error occurs for methane. For the 96 compounds that were studied, the average errors in ionization potentials are 0.72, 0.53, 0.59, 0.65, 0.56, and 1.20 eV for MNDO, AM1, PM3, PDDG/MNDO, PDDG/PM3, and NO-MNDO. However, in plots of the experimental and SMO results, as in Figure 1, the correlation coefficients (r^2) for the ionization potentials are 0.75, 0.86, 0.81, 0.82, 0.82, and 0.80 from MNDO, AM1, PM3, PDDG/MNDO, PDDG/PM3, and NO-MNDO, and rms errors from the linear fits are 0.74, 0.55, 0.64, 0.62, 0.62, and 0.66 eV, respectively.

For gas-phase dipole moments, 47 molecules were considered. The average errors are 0.29, 0.22, 0.25, 0.20, 0.23, and 0.31 D from MNDO, AM1, PM3, PDDG/MNDO, PDDG/PM3, and NO-MNDO. For the correlations of the experimental and computed values, the corresponding r^2 values are 0.88, 0.90, 0.91, 0.93, 0.92, and 0.88, and the rms errors are 0.38, 0.33, 0.33, 0.30, 0.31, and 0.37 D, respectively.

5. Conclusions

The effect of introduction of the overlap matrix in the secular determinant has been evaluated starting from the MNDO method. The implementation featured parametrization for molecules containing C, H, N, and O atoms, and the resultant nonorthogonal method was designated NO-MNDO. Testing included computation of a large number and variety of energetic quantities, ionization potentials, and dipole moments. Any study of this type may be incomplete since additional testing, e.g., for ion energetics or hydrogen bonding, could be performed and because the optimal parameter sets may not have been found, as suspected here for nitrogen with NO-MNDO. The present results indicate that the NO-MNDO modification coupled with optimization of the atomic energies, *eisol*, provides significantly improved

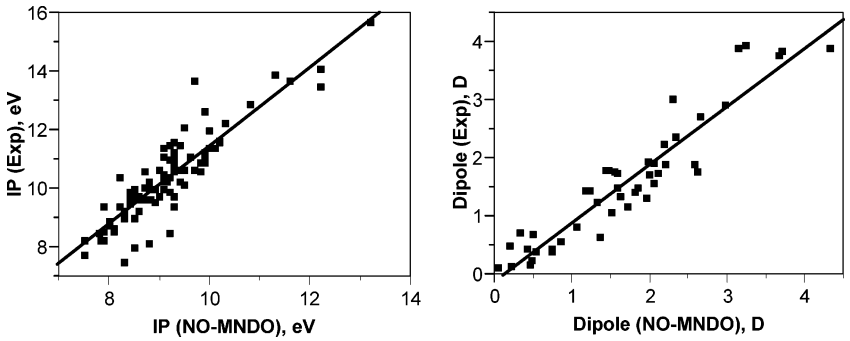


Figure 1. Correlation of experimental ionization potentials (left) and gas-phase dipole moments (right) with results from NO-MNDO calculations.

energetic results over those from MNDO. Overall, the accuracy of NO-MNDO is very similar to that of the AM1 method, which utilizes more than 30 additional, optimized parameters. Therefore, it is apparent that one can devise an MNDO variant that has similar quality as AM1 but does not require the addition of the AM1 Gaussians to the core repulsion formula. Notable improvements for NO-MNDO over MNDO are obtained for rotational barriers about single bonds and for the barriers for hydrogen-atom transfer reactions. The characteristic branching errors for isomers from MNDO and AM1 were also relieved. It is reiterated that the present work was not carried out to introduce a new SMO method; its sole purpose was to test the impact of including the overlap matrix in the secular equations for an MNDO-based method. The associated improvements are significant enough to warrant consideration of the methodological change in the development of future semiempirical MO methods.

Acknowledgment. The authors wish to thank Dr. Cristiano R. W. Guimarães for helpful discussion and the National Science Foundation (CHE-0446920) for financial support.

Supporting Information Available: Complete listing of all computed and experimental data. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Zerner, M. C. In *Semiempirical Molecular Orbital Methods*; Lipkowitz, K. B., Boyd, D. B., Eds.; VCH Publishers: New York, 1991; Vol. 2, p 313.

(2) Stewart, J. J. P. In *Semiempirical Molecular Orbital Methods*; Lipkowitz, K. B., Boyd, D. B., Eds.; VCH Publishers: New York, 1991; Vol. 1, p 70.

(3) Pople, J. A.; Beveridge, D. L. In *Approximate Molecular Orbital Theory*; McGraw-Hill: New York, 1970.

(4) Dewar, M. J. S.; Thiel, W. *J. Am. Chem. Soc.* **1977**, *99*, 4899.

(5) Dewar, M. J. S.; Thiel, W. *J. Am. Chem. Soc.* **1977**, *99*, 4907.

(6) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. *J. Am. Chem. Soc.* **1985**, *107*, 3902.

(7) Stewart, J. J. P. *J. Comput. Chem.* **1989**, *10*, 209.

(8) Stewart, J. J. P. *J. Comput. Chem.* **1989**, *10*, 221.

(9) Thiel, W.; Voityuk, A. A. *J. Phys. Chem.* **1996**, *100*, 616.

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

(11) Bartlett, R. J.; Watts, J. D.; Kucharski, S. A.; Noga, J. *Chem. Phys. Lett.* **1990**, *165*, 513.

(12) Curtiss, L. A.; Raghavachari, K.; Trucks, G. W.; Pople, J. A. *J. Chem. Phys.* **1991**, *94*, 7221.

(13) Curtiss, L. A.; Raghavachari, K.; Trucks, G. W.; Pople, J. A. *J. Chem. Phys.* **1997**, *106*, 1063.

(14) Dewar, M. J. S.; Dieter, K. M. *J. Am. Chem. Soc.* **1986**, *108*, 8.

(15) Burk, P.; Herodes, K.; Koppel, I.; Koppel, I. *Int. J. Quantum Chem. Symp.* **1993**, *27*, 633.

(16) Repasky, M. P.; Chandrasekhar, J.; Jorgensen, W. L. *J. Comput. Chem.* **2002**, *23*, 1601.

(17) (a) Weber, W.; Thiel, W. *Theor. Chem. Acc.* **2000**, *103*, 495. (b) Mohle, K.; Hofmann, H.-J.; Thiel, W. *J. Comput. Chem.* **2001**, *22*, 509. (c) Kolb, M.; Thiel, W. *J. Comput. Chem.* **1993**, *14*, 775.

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

(19) Hall, G. G. *Proc. R. Soc. London, Ser. A* **1951**, *205*, 541.

(20) Pople, J. A.; Santry, D. P.; Segal, G. A. *J. Chem. Phys.* **1965**, *43*, S129.

(21) Hoffmann, R. *J. Chem. Phys.* **1963**, *39*, 1397.

(22) Jorgensen, W. L.; Borden, W. T. *J. Am. Chem. Soc.* **1973**, *95*, 6649.

(23) (a) Tubert-Brohman, I.; Guimarães, C. R. W.; Reparsky, M. P.; Jorgensen, W. L. *J. Comput. Chem.* **2004**, *25*, 138. (b) Tubert-Brohman, I.; Guimarães, C. R. W.; Jorgensen, W. L. *J. Chem. Theory Comput.* In press.

(24) Stewart, J. J. P. *J. Comput.-Aided Mol. Des.* **1990**, *4*, 1.

(25) Stewart, J. J. P. *J. Mol. Model.* **2004**, *10*, 6.

(26) Guner, V.; Khuong, K. S.; Leach, A. G.; Lee, P. S.; Bartberger, M. D.; Houk, K. N. *J. Phys. Chem.* **2003**, *107*, 11445.

(27) Ess, H. D.; Houk, K. N., personal communication.

(28) Dybala-Defratyka, A.; Paneth, P.; Pu, J.; Truhlar, D. G. *J. Phys. Chem.* **2004**, *108*, 2475−2486.

(29) Herrebout, W. A.; van der Veken, B. J.; Wang, A.; Durig, J. R. *J. Phys. Chem.* **1995**, *99*, 578.

(30) (a) Murcko, M. A.; Castejon, H.; Wiberg, K. B. *J. Phys. Chem.* **1996**, *100*, 16162. (b) Allinger, N. L.; Fermann, J. T.; Allen, W. D.; Schaefer, H. F., III *J. Chem. Phys.* **1997**, *106*, 5143.

(31) Weiss, S.; Leroi, G. *J. Chem. Phys.* **1968**, *48*, 962.

(32) Csaszár, A. G.; Allen, W. D.; Schaefer, H. F., III *J. Chem. Phys.* **1998**, *108*, 9751.

(33) Moazzen-Ahmadi, N.; Gush, H. P.; Halpren, M.; Jagannath, H.; Leung, A.; Ozier, I. *J. Chem. Phys.* **1988**, *88*, 563.

(34) Fantoni, R.; van Hellvoort, K.; Knippers, W.; Reuss, J. *Chem. Phys.* **1986**, *110*, 1.

(35) Booth, H.; Everett, J. R. *J. Chem. Soc., Perkin Trans.* **1980**, *2*, 255.

(36) Abraham, R. J.; Ribeiro, D. S. *J. Chem. Soc., Perkin Trans.* **2001**, *2*, 302.

(37) Wiberg, K. B.; Martin, E. *J. Am. Chem. Soc.* **1985**, *107*, 5035.

(38) Murphy, R. B.; Beachy, M. D.; Friesner, R. A.; Ringnalda, M. N. *J. Chem. Phys.* **1995**, *103*, 1481.

CT050174C

JCTC Journal of Chemical Theory and Computation

Improved Efficiency of Replica Exchange Simulations through Use of a Hybrid Explicit/Implicit Solvation Model

Asim Okur,[†] Lauren Wickstrom,[‡] Melinda Layten,[§] Raphäel Geney,[†] Kun Song,[†] Viktor Hornak,[||] and Carlos Simmerling*[,†,‡,||,⊥]

Department of Chemistry, Graduate Program in Biochemistry and Structural Biology, Graduate Program in Molecular and Cellular Biology, and Center for Structural Biology, Stony Brook University, Stony Brook, New York 11794, and Computational Science Center, Brookhaven National Laboratory, Upton, New York 11973

Received August 5, 2005

Abstract: The use of parallel tempering or replica exchange molecular dynamics (REMD) simulations has facilitated the exploration of free energy landscapes for complex molecular systems, but application to large systems is hampered by the scaling of the number of required replicas with increasing system size. Use of continuum solvent models reduces system size and replica requirements, but these have been shown to provide poor results in many cases, including overstabilization of ion pairs and secondary structure bias. Hybrid explicit/continuum solvent models can overcome some of these problems through an explicit representation of water molecules in the first solvation shells, but these methods typically require restraints on the solvent molecules and show artifacts in water properties due to the solvation interface. We propose an REMD variant in which the simulations are performed with a fully explicit solvent, but the calculation of exchange probability is carried out using a hybrid model, with the solvation shells calculated on the fly during the fully solvated simulation. The resulting reduction in the perceived system size in the REMD exchange calculation provides a dramatic decrease in the computational cost of REMD, while maintaining a very good agreement with results obtained from the standard explicit solvent REMD. We applied several standard and hybrid REMD methods with different solvent models to alanine polymers of 1, 3, and 10 residues, obtaining ensembles that were essentially independent of the initial conformation, even with explicit solvation. Use of only a continuum model without a shell of explicit water provided poor results for Ala_3 and Ala_{10}, with a significant bias in favor of the α-helix. Likewise, using only the solvation shells and no continuum model resulted in ensembles that differed significantly from the standard explicit solvent data. Ensembles obtained from hybrid REMD are in very close agreement with explicit solvent data, predominantly populating polyproline II conformations. Inclusion of a second shell of explicit solvent was found to be unnecessary for these peptides.

Introduction

The potential energy surfaces of biological systems have long been recognized to be rugged, hindering conformational transitions between various local minima. This sampling problem can preclude success even when a sufficiently accurate Hamiltonian of the system is used in the simulations. Thus, a significant effort has been put into devising efficient simulation strategies that locate low-energy minima for these

* Corresponding author e-mail: carlos.simmerling@stonybrook.edu.
[†] Department of Chemistry, Stony Brook University.
[‡] Graduate Program in Biochemistry and Structural Biology, Stony Brook University.
[§] Graduate Program in Molecular and Cellular Biology, Stony Brook University.
[||] Center for Structural Biology, Stony Brook University.
[⊥] Brookhaven National Laboratory.

10.1021/ct050196z CCC: $33.50
Published on Web 02/04/2006

complex systems. Conformational sampling was recently reviewed[1] and is also the subject of a recent special journal issue.[2]

One approach that has seen a recent increase in the use of biomolecular simulation is the replica exchange method.[3−5] In replica exchange molecular dynamics (REMD)[6] (also called parallel tempering[3]), a series of molecular dynamics simulations (replicas) are performed for the system of interest. In the original form of REMD, each replica is an independent realization of the system, coupled to a heat bath at a different temperature. The temperatures of the replicas span a range from low values of interest (such as 280 K or 300 K) up to high values (such as 600 K) at which the system can rapidly overcome potential energy barriers that would otherwise impede conformational transitions on the time scale simulated.

At intervals during the otherwise standard simulations, conformations of the system being sampled at different temperatures are exchanged based on a Metropolis-type criterion[7] that considers the probability of sampling each conformation at the alternate temperature (described in more detail in Methods). In this manner, REMD is hampered to a lesser degree by the local minima problem, since simulations at low temperature can escape kinetic traps by "jumping" directly to alternate minima being sampled at higher temperatures. Likewise, the structures sampled at high temperatures can anneal by being transferred to successively lower temperatures. Moreover, the transition probability is constructed such that the canonical ensemble properties are maintained during each simulation, thus providing potentially useful information about conformational probabilities as a function of temperature. Due to these advantages, REMD has been applied to studies of peptide and small protein folding.[3,6,8−16]

For large systems, however, REMD becomes intractable since the number of replicas needed to span a given temperature range increases with the square root of the number of degrees of freedom in the system.[17−20] Several promising techniques have been proposed[19,21−23] to deal with this apparent disadvantage to REM.

The method chosen to treat solvent effects can have a direct impact on the system size and thus the computational requirement of employing REMD. Explicit representation of solvent molecules significantly increases the number of atoms in the simulated system, particularly when the solvent box is made large enough to enclose unfolded conformations of peptides and proteins. The growth in system size results in the need for many more replicas to span the same temperature range. This increase in computational cost is in addition to that added by the need to calculate forces and integrate equations of motion for the explicit solvent molecules.

Continuum solvent models such as the semianalytical Generalized Born (GB) model[24] estimate the free energy of solvation of the solute based on coordinates of the solute atoms. The neglect of explicit solvent molecules can significantly reduce the computational cost of evaluating energies and forces for the system, but a larger effect with REMD can arise from the reduction in the number of replicas due to the fewer degrees of freedom. This factor can determine whether REMD is a practical approach to model the system. For example, in the 10-residue peptide model presented below, 40 replicas are needed when the solvent is included explicitly, while only 8 are sufficient for the same peptide with a continuum solvent model. Larger systems would be expected to show even greater differences; the number of peptide atoms increases approximately linearly with sequence length, while the volume of a sphere (and thus the number of solvent atoms) needed to enclose extended conformations increases with the peptide length to the third power. Thus one can roughly estimate that the difference in number of replicas required for explicit vs continuum solvation of a system will increase with the number of solute degrees of freedom to the $^3/_2$ power.

Continuum solvent models are thus an attractive approach to enabling the study of larger systems with REMD. Among the various models that have been developed, the GB approach is commonly used with molecular dynamics due to its computational efficiency, permitting use at each time step. However, these models can also have significant limitations. Since the atomic detail of the solvent is not considered, modeling specific effects of structured water molecules can be challenging. In the case of protein and peptide folding, it appears likely that the current generation of GB models do not have as good a balance between protein−protein and protein−solvent interactions as do the more widely tested explicit solvent models.[25,26] More particularly, it has been reported[12,26−28] that ion pairs were frequently too stable in the GB implicit water model, causing salt bridged conformations to be oversampled in MD simulations, thus altering the thermodynamics and kinetics of folding for small peptides. A clear illustration was given by Zhou and Berne[26] who sampled the C-terminal β-hairpin of protein G (GB1) with both a surface-GB (SGB)[29] continuum model and an explicit solvent. The lowest free energy state with SGB was significantly different from the lowest free energy state in the explicit solvent, with incorrect salt bridges formed at the core of the peptide, in place of hydrophobic contacts. Zhou extended this study on GB1 by examining several force field-GB model combinations, with all GB models tested showing erroneous salt-bridges.[27]

The more rigorous models based on Poisson−Boltzmann (PB) equations are generally considered to be more accurate. Historically, the increased cost of evaluating solvation free energy with these methods results in their use primarily to postprocess a small number of conformations, or snapshots sampled during an MD simulation in the explicit solvent.[30] However, some researchers have reported using PB as a solvent model for molecular dynamics simulation.[31,32] PB approaches do not necessarily overcome the difficulty of modeling nonbulk effects in the first solvation shells.

To benefit from the efficiency of implicit solvents while incorporating these first shell effects, several hybrid explicit/implicit models have been proposed. These typically employ the explicit solvent only for the first 1−2 solvation shells of the solute, often surrounded by a continuum representation of various types.[33−45] However, these methods have draw-

backs in that the explicit water typically must be restrained to remain close to the solute to avoid diffusion into the "bulk" continuum. These restraints as well as the boundary effects at the explicit/implicit interface can have a dramatic effect on solute behavior. In a recent implementation, Lee et al. employed a hybrid TIP3P/GB solvation model with excellent results,[41] but they pointed out drawbacks typical for these models, such as the need for a fixed solute volume and shape for the solvation cavity, preventing large-scale conformational changes of the type that is necessary for detailed analysis of conformational ensembles using enhanced sampling techniques such as REMD. In addition, they demonstrated that solvent properties such as radial density and dipole distributions showed significant artifacts due to boundary effects.

Recognizing that the main difficulty in applying REMD with the explicit solvent lies in the number of simulations required, rather than just the complexity of each simulation, we propose a new approach in which each replica is simulated in the explicit solvent using standard methods such as periodic boundary conditions and inclusion of long-range electrostatic interactions. However, the calculation of exchange probabilities (which determines the temperature spacing and thus the number of replicas) is handled differently. Only a subset of closest water molecules is retained, while the remainder is *temporarily* replaced by a continuum representation. The energy is calculated using the hybrid model, and the exchange probability is determined. The original solvent coordinates are then restored, and the simulation proceeds as a continuous trajectory with fully explicit solvation. This way the perceived system size for evaluation of exchange probability is dramatically reduced and fewer replicas are needed.

An important difference from the existing hybrid models is that our system is fully solvated throughout the entire simulation, and thus the distribution functions and solvent properties should not be affected by the use of the hybrid model in the exchange calculation. In addition, no restraints of any type are needed for the solvent, and the solute shape and volume may change since the solvation shells are generated for each replica on the fly at every exchange calculation. Nearly no computational overhead is involved since the calculation is performed infrequently as compared to the normal force evaluations. Thus the hybrid REMD approach can employ more accurate continuum models that are too computationally demanding for use in each time step of a standard molecular dynamics simulation.

In this study we have tested the hybrid REMD method on varying lengths of polyalanine peptides (dipeptide, tetrapeptide, and Ala_{10}). Many helical design studies have used polyalanines with charged residues,[46−48] N-capping,[49] and C-capping interactions[50] to solubilize the peptides and stabilize helical structure. Recently, experimental studies with CD, NMR, and UV resonance Raman have been able to characterize a primarily polyproline type II (P_{II}) structure in short polyalanines[51−53] and in the denatured state of longer alanine peptides.[54] MD simulations of polyalanines have further substantiated these experimental observations.[38,55] The quality of the solvent model is expected to be critically important since it has been proposed that specific solvation of backbone amide groups plays a key role in the stabilization of P_{II} conformations.[55,56]

For each peptide we first obtained conformation ensembles using standard REMD in explicit solvent. We used these data as a reference in order to remove the influence of the protein force field parameters from this study of solvation models. For each sequence, two sets of REMD simulations in the explicit solvent were run with different initial conformations until convergence was indicated by reasonable agreement between the data sets. For example, the populations of conformation clusters in the two Ala_{10} runs in the TIP3P solvent were highly correlated (R^2=0.974), demonstrating high similarity not only in the types of structures sampled in these two simulations but also in their probability in these independently generated ensembles. This level of convergence gives us confidence that the differences we observe between the various solvent models are predominantly due to solvation effects and not poorly converged ensembles with large uncertainties in the resulting data.

We then employed pure GB REMD simulation using both models available in Amber (GB^{HCT} [57] and GB^{OBC} [58,59]) as well as the hybrid REMD approach using the same GB models. We also performed REMD where only the first 1 or 2 solvation shells were retained for the exchange calculations (without a continuum model). Comparison of these results to each other and to the standard explicit solvent REMD results provides insight into the performance of the GB models, the improvement obtained by retaining the first solvation shell in the calculation of exchange probability (the hybrid model), and the need for the reaction field surrounding the solvation shells.

We compared ensemble distributions of properties such as chain end-to-end distance, backbone ϕ/ψ free energy maps, and cluster populations among the methods. While all of the solvation models provided similar results for alanine dipeptide, the GB models failed to reproduce the TIP3P ensemble data for Ala_3 and Ala_{10} even at a qualitative level, providing ensembles that were dominated by α-helical conformations. Simulations using hybrid REMD using GB^{OBC} and only a single shell of explicit water were in good accord with the reference simulations, with a high degree of similarity between structure populations (R^2=0.93), with lack of significant α-helix, and a strong preference for P_{II} conformation. This agreement was obtained despite a significant reduction in computational cost; for Ala_{10}, 40 replicas were used for standard REMD in TIP3P, while only 8 were needed for pure GB or hybrid GB/TIP3P REMD.

Methods

Replica Exchange Molecular Dynamics (REMD). We briefly summarize the key aspects of REMD as they relate to the present study. In standard Parallel Tempering or Replica Exchange Molecular Dynamics,[3,6] the simulated system consists of M noninteracting copies (replicas) at M different temperatures. The positions, momenta, and temperature for each replica are denoted by $\{q^{[i]}, p^{[i]}, T_m\}$, $i =$

1,...,*M*; *m* = 1,..., *M*. The equilibrium probability for this generalized ensemble is

$$W(p^{[i]},q^{[i]},T_m) = \exp\left\{-\sum_{i=1}^{M}\frac{1}{k_B T_m}H(p^{[i]},q^{[i]})\right\} \quad (1)$$

where the Hamiltonian $H(p^{[i]},q^{[i]})$ is the sum of kinetic energy $K(p^{[i]})$ and potential energy $E(q^{[i]})$. For convenience we denote $\{p^{[i]},q^{[i]}\}$ at temperature T_m by $x_m^{[i]}$ and further define $X = \{x_1^{[i(1)]}, ..., x_M^{[i(M)]}\}$ as one state of the generalized ensemble. We now consider exchanging a pair of replicas. Suppose we exchange replicas *i* and *j*, which are at temperatures T_m and T_n, respectively,

$$X = \{ ...;x_m^{[i]};...;x_n^{[j]};...\} \rightarrow X' = \{ ...;x_m^{[j]};...;x_n^{[i]};...\} \quad (2)$$

To maintain a detailed balance of the generalized system, microscopic reversibility has to be satisfied, thus giving

$$W(X)\rho(X\rightarrow X') = W(X')\rho(X'\rightarrow X) \quad (3)$$

where $\rho(X\rightarrow X')$ is the exchange probability between two states *X* and *X*′. With the canonical ensemble, the potential energy *E* rather than total Hamiltonian *H* will be used simply because the momentum can be integrated out. Inserting eq 1 into eq 3, the following equation for the Metropolis exchange probability is obtained:

$$\rho = \min\left(1, \exp\left\{\left(\frac{1}{k_B T_m} - \frac{1}{k_B T_n}\right)(E(q^{[i]}) - E(q^{[j]}))\right\}\right) \quad (4)$$

In practice, several replicas at different temperatures are simulated simultaneously and independently for a chosen number of MD steps. Exchange between a pair of replicas is then attempted with a probability of success calculated from eq 4. If the exchange is accepted, the bath temperatures of these replicas will be swapped, and the velocities will be scaled accordingly. Otherwise, if the exchange is rejected, each replica will continue on its current trajectory with the same thermostat temperature.

As we described above, one of the major limitations of REM is that the number of replicas needed to span a temperature range grows proportionally to the square root of number of degrees of freedom in the simulated system. While a more rigorous analysis of the acceptance probability in REM trials has been given recently using a Gaussian energy distribution model,[20,60] one can also approximate from eq 4 that the overall exchange probability P_{acc} is proportional to $\exp(-\Delta T^2/T^2)$, which implies that a greater acceptance ratio requires a smaller temperature gap ΔT or a more dense temperature distribution to reach. On the other hand, ΔT should be as large as possible so as to span a wide temperature range with a small number of replicas. The relationship can be estimated through consideration of potential energy fluctuations of two replicas sampling at the target temperature T_n and T_{n-1} (Figure 1). The instantaneous energy fluctuation δE in a given simulation at temperature *T* scales as $\sqrt{f}T$, and the average energy gap ΔE between two neighboring replicas is proportional to $f\Delta T$, where *f* is the number of degrees of freedom and $\Delta T = T_n - T_{n-1}$. Obtaining a reasonable acceptance ratio relies on keeping the replica energy gap comparable to the energy fluctuations, thus $\Delta E/\delta E$ should be near unity. Since $\Delta E/\delta E$ is proportional to $\Delta T\sqrt{f}/T$, the acceptable temperature gap between neighboring replicas therefore decreases with larger systems as $\Delta T \sim 1/\sqrt{f}$, and more simultaneous simulations are needed to cover the desired temperature range.

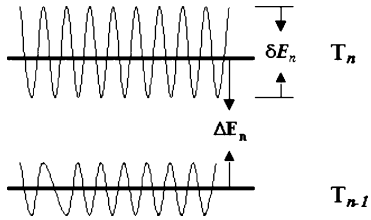


Figure 1. Schematic diagram illustrating the energy fluctuations for simulations at two temperatures for neighboring replicas. To obtain high exchange probabilities, the energy fluctuations δE in each simulation should be of comparable magnitude to the mean energy difference ΔE.

Model Systems and Simulation Details. We simulated three polyalanine sequences: alanine dipeptide (Ala_1), alanine tetrapeptide (Ala_3), and polyalanine (Ala_{10}), all with acetylated and amidated N- and C-termini, respectively. All simulations employed the Amber ff99 force field,[61,62] with modifications[63] to reduce α-helical bias. Explicit solvent and hybrid REMD used the TIP3P water model.[64] The standard REMD simulations in explicit solvent and in pure GB were run using our REMD implementation as distributed in Amber (version 8).[65] The hybrid solvent REMD calculations were performed with a locally modified version of Amber 8. All bonds involving hydrogen were constrained in length using SHAKE.[66] The time step was 2 fs. Temperatures were maintained using weak coupling[67] to a bath with a time constant of 0.5 ps^{-1}.

Secondary structure basin populations for central residues were calculated based on ϕ/ψ dihedral angle pairs. The dihedral angle ranges defining for those regions are provided in Table S1. The solvent accessible surface areas (SASA) for simulated peptides were calculated using the gbsa = 2 option in AMBER. The end-to-end distances for Ala_{10} were calculated between Cα atoms of Ala2 and Ala9 (omitting terminal residues) using the ptraj module of Amber. Cluster analysis for Ala_{10} was performed using moil-view,[68] using backbone RMSD for Ala2−9 and a similarity cutoff of 2.5 Å.

Explicit Solvent REMD. The Ala_{10} peptide in α-helical conformation was solvated in a truncated octahedral box using 983 TIP3P water molecules for a total of 3058 atoms. The system was equilibrated at 300 K for 50 ps with harmonic positional restraints on solute atoms, followed by minimizations with gradually reduced solute positional restraints and three 5 ps MD simulations with gradually reduced restraints at 300 K. Long-range electrostatic interactions were calculated using PME.[69] Simulations were run in the NVT ensemble.

Forty replicas were used at temperatures ranging from 267 K to 571 K, which were optimized to give a uniform exchange acceptance ratio of ∼30%. Exchange between neighboring temperatures was attempted every 1 ps, and each

REMD simulation was run for 50 000 exchange attempts (50 ns). The first 5 ns of each simulation was discarded to remove the initial structure bias.

To provide a stringent test of data convergence for greater conformational diversity expected for Ala_{10}, two sets of REMD simulations were performed, starting from different initial conformations. In one set, all replicas were started from a fully α-helical conformation; in the other an extended conformation was employed. In the case of Ala_1 and Ala_3, lower bounds for uncertainty were estimated by separating the full simulation data into halves and reporting the difference between values calculated for each half.

A similar procedure was used for Ala_1 and Ala_3. Ala_1 was solvated in a truncated octahedral box using 341 TIP3P water molecules. Ala_3 required 595 water molecules. For both systems the same equilibration procedure as used for Ala_{10} was employed. To cover the same temperature range 20 replicas for Ala_1 and 26 replicas for Ala_3 were needed. Both systems were simulated for ∼40 000 exchanges, and the first 5000 exchange attempts were discarded as equilibration.

Implicit Solvent REMD. Solvent effects were calculated through the use of two Generalized Born implementations in Amber (GB^{HCT} and GB^{OBC} (note that GB^{OBC} is model 2 in ref 59)). Two sets of intrinsic Born radii were used, both adopted from Bondi[70] with modification of hydrogen.[71] Unless otherwise noted, the GB^{HCT} model was used with the mbondi radii, and the GB^{OBC} model was employed with mbondi2 radii (as recommended in Amber). Scaling factors were taken from the TINKER modeling package.[72] No cutoff on nonbonded interactions was used. All other simulation parameters were the same as used in explicit solvent.

For Ala_{10}, the use of the continuum solvent model resulted in a total of 109 atoms considered explicitly in the simulations (∼28 times fewer than in the explicitly solvated system). The much smaller system size permitted the use of 8 replicas to cover the same temperature range that required 40 replicas in the explicit solvent, while obtaining the same 30% exchange acceptance probability. Exchanges were attempted every 1 ps, and the REMD simulation was run for 50 000 exchange attempts (50 ns). Simulations were initiated with the same two initial conformation ensembles as were used for the explicit solvent REMD calculations, with comparison of the two runs providing a lower bound for the uncertainty in resulting data. For Ala_1 and Ala_3 the same approach was used, with 4 replicas used to cover the temperature space for each system. Simulations were run for 50 000 exchange attempts, and the first 5000 exchanges were discarded.

Hybrid Solvent REMD. All simulation parameters in the hybrid solvent REMD simulations were the same as those employed for standard REMD in the explicit solvent, with the exception that the number of replicas (8 for Ala_1, Ala_3, and Ala_{10}) and the target temperatures were the same as those used for the pure GB REMD simulations for Ala_{10}. It is important to note that the hybrid solvent model was used *only* for calculation of exchange probability; the simulations themselves were performed on fully solvated systems with truncated octahedral periodic boundary conditions and PME for the calculation of long-range electrostatic interactions.

We determined the number of water molecules to retain in the hybrid model based on analysis of the number of waters in the first solvation shell of Ala_{10} in the ensemble of structures sampled in the standard REMD explicit solvent simulations. We found that 100 water molecules were sufficient even for the most extended conformations (data not shown). Thus this number was used for all replicas and all exchanges. For Ala_1, 30 water molecules were enough to incorporate the first solvation shell and 60 water molecules for the first and second solvation shells. These numbers increase to 50 waters and 100 waters for the first solvation shell and the first and second solvation shells of Ala_3, respectively. Ala_1 and Ala_3 hybrid simulations were run for ∼30 000 exchanges, and the first 5000 were discarded.

At each exchange step, the distance between the oxygen atom of each water molecule and all solute atoms was calculated. Water molecules were then sorted by their closest solute distance, and all water molecules except the X with the shortest solvent−solute distances were temporarily discarded (where X is the number of waters retained in each system, as described above). The energy of this smaller system was then recalculated using only these close waters and the GB solvent model. This energy was used to calculate the exchange probability, and then all waters were restored to their original positions and the simulations were continued (Figure 2). In this manner the simulations using the hybrid solvent model were continuous simulations with fully solvated PBC/PME, and the hybrid model was used only for the calculation of exchange probabilities.

Results and Discussion

Comparison of Exchange Efficiency for Hybrid and Standard REMD in Ala_{10}. Even though REMD has become a useful tool to improve conformational sampling, REMD simulations are highly computationally expensive, particularly when the solvent is treated explicitly. The increase in cost arises not only from the additional effort involved in calculating forces in a given simulation but also from the increase in the number of simulations (replicas) needed to span a particular temperature range. This increase is due to the much larger number of degrees of freedom present in the explicitly solvated system as compared to that in continuum solvent models. In the case of Ala_{10}, our largest model system, the number of replicas needed to span the range of 267 K to 571 K increases from 8 to 40 when switching from implicit to explicit solvation.

We evaluated the utility of the hybrid solvent model during the calculation of the exchange probability on several levels, using Ala_{10} as its size is most relevant to the larger systems that would benefit most from this method. First, we validated that fewer replicas were needed to obtain efficient exchange with the hybrid model as compared to the number required when retaining the full periodic box of explicit water molecules during the exchange probability calculation (eq 4). Efficient exchanges were obtained with the hybrid model even when using the same number of replicas as was needed for the pure continuum solvent REMD simulations. Next, we evaluated whether the use of the hybrid model affected the data obtained from the simulations, with particular

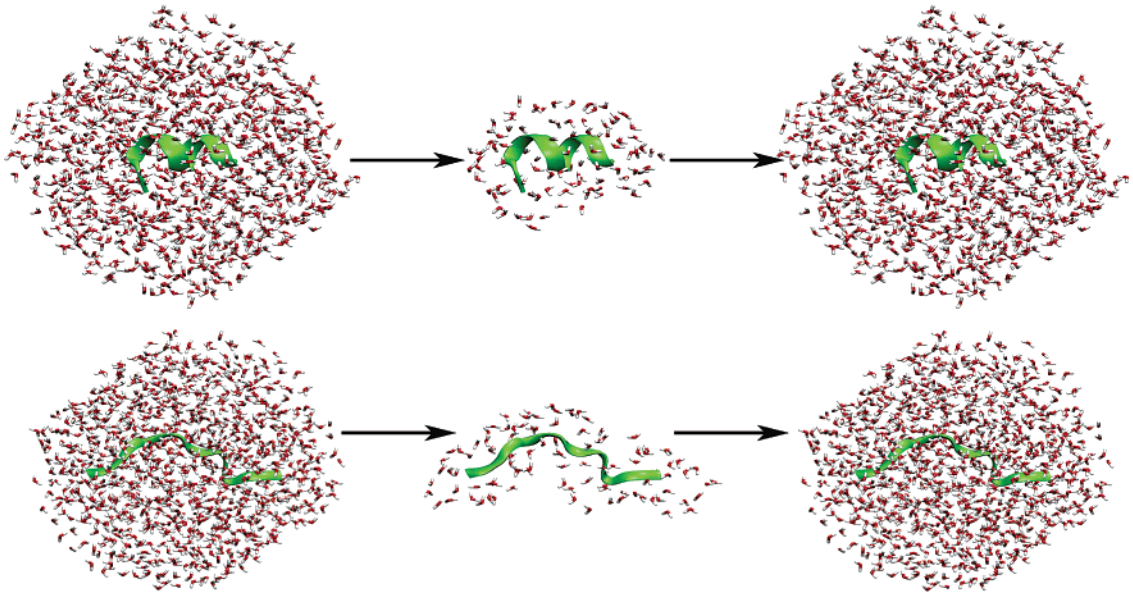


Figure 2. Schematic description of hybrid solvent REMD. The fully solvated Ala_{10} (with truncated octahedral boundary conditions) is simulated between exchanges (left). The exchange energy is calculated by retaining only the closest 100 waters (center), with bulk solvent properties calculated using the GB solvation model. After the exchange calculation the explicit solvent is restored, and the dynamics continues under periodic boundary conditions. This approach allows on the fly calculation of the solvation shell, whose shape adjusts automatically to the solute conformation (top: α-helical structure, bottom: extended structure). As a result, many fewer replica simulations are required.

emphasis on the conformational distributions sampled by the model peptides. These distributions were also compared to those obtained for REMD with only the continuum solvent model.

An important benefit of REMD is the ability to obtain improved sampling at low temperatures of interest by exchanging conformations with higher temperature simulations that have less likelihood to become kinetically trapped. As described in Methods, the probability of the successful exchange of conformations between two temperatures depends on the overlap in potential energy distributions at those temperatures. Figure 3 shows the potential energy distributions for each temperature for sets of simulations with explicit solvent (A) and those with GB (B) between 267 K and 571 K. The graph illustrates why fewer replicas are required for the GB model; the energy range spanned is smaller for the smaller system, and fewer replicas are still able to achieve the required overlap. In contrast, when the explicit solvent model is used with only the 8 replica temperatures that are successful with GB, no significant overlap in the distributions is observed (Figure 3C).

Based on Figure 3, exchanges between replicas at neighboring temperatures are expected to occur with a high probability when using 40 replicas in explicit solvent or 8 replicas with GB. No exchanges are expected for the explicit solvent with only 8 replicas. Figure 4 shows the temperature histories of the first 2 replicas in the same explicit solvent and GB REMD simulations as were shown in Figure 3. As expected, the replicas visited all available temperatures during the run (the other replicas showed similar behavior and are not shown for clarity). However, the explicit solvent REMD with only 8 replicas showed *no* exchanges even after 25 000 attempts (25 ns simulation), and all replicas remained at their initial temperatures. This REMD simulation is identical to

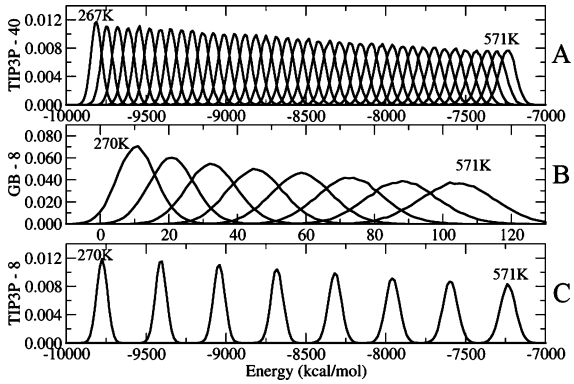


Figure 3. Potential energy distributions for Ala_{10} simulations over a range of temperatures using (A) explicit solvent REMD with 40 replicas, (B) GB REMD with 8 replicas, and (C) explicit solvent REMD with 8 replicas using the same temperature distribution as GB REMD. GB simulations involve fewer degrees of freedom and are able to span the energy range with fewer replicas. In contrast, no overlap is obtained when using explicit solvent with the same replica and temperature selection as GB. This implies that no exchanges would be permitted, and the benefits of REMD would be lost.

8 standard MD simulations at different temperatures, and therefore no sampling improvement is obtained. Thus, in order for replicas to sample a range of temperatures, more replicas (and significantly more computational resources) are required for simulations in the explicit solvent. Reducing this requirement while maintaining fully explicitly solvated simulations is the goal of our hybrid model.

These exchange efficiencies are all consistent with previously reported REMD simulations and the known scaling with system size of the number of replicas required for efficient exchange. In our case these data provide an

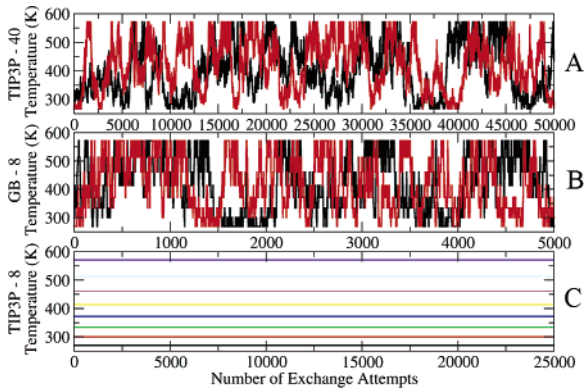


Figure 4. Temperature histories for Ala_{10} replicas using (A) explicit solvent with 40 replicas, (B) GB with 8 replicas, and (C) explicit solvent with 8 replicas. For clarity only the first two replicas for A and B and only the first 5000 exchanges of B are shown. Consistent with the potential energy distributions shown in Figure 3, exchanges are only obtained when sufficient overlap in potential energy distributions is present. If too few replicas are used (C), the result is a series of standard MD simulations.

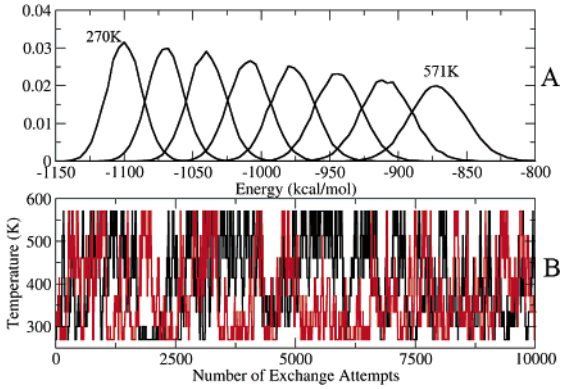


Figure 5. Potential energy distributions (A) and temperature histories of 2 Ala_{10} replicas (B) using 8 replicas in periodic boxes with fully explicit solvent but with the hybrid solvent model for calculation of exchange probability. Use of the hybrid model gives overlap between neighboring temperatures and allows replicas to span a range of temperatures, in sharp contrast to the total lack of exchanges for the *same* simulated system with standard REMD (Figures 3C and 4C). For clarity only the first 10 000 exchanges are plotted, and only 2 replicas are shown in the lower figure.

important context for evaluation of the use of hybrid solvation models during the calculation of exchange probability. We performed REMD simulations using the same explicitly solvated system as shown above, but with only the 8 replicas/temperatures that gave an efficient exchange with pure GB solvation. With standard REMD, this system showed no overlap in potential energy distributions and was unable to generate any successful exchanges (Figure 4C). We employed the hybrid solvent model only for calculation of the exchange probability (eq 4) for this fully explicit solvent system. The distributions of the potential energies for the different temperatures during 10 000 exchange attempts (10 ns) are shown in Figure 5. Use of the hybrid solvent model permits the simulations to achieve nearly the same level of energy distribution overlap as we obtained for the pure GB model. Consistent with this observation, multiple exchanges are observed despite the relatively small number of replicas employed. The replicas are able to traverse the entire temperature range on the nanosecond time scale. It is interesting to note that this is more rapid than seen for the standard REMD explicit solvent run, most likely due to the larger temperature step taken with each successful exchange with the hybrid solvent model (due to larger ΔT between neighboring replicas). The standard REMD run requires more exchanges to traverse the same total temperature range. This suggests that the hybrid calculation may have additional advantages beyond simply reducing the number of replicas as compared to the standard REMD; however, such an analysis is outside the scope of the present article.

Analysis of Conformational Sampling in Hybrid and Standard REMD. After establishing the ability of the hybrid REMD model to reduce the number of replicas required to obtain efficient exchanges, we examined the ability of the hybrid approach to reproduce ensemble data obtained with standard REMD in the explicit solvent. We also investigated whether the reaction field beyond the solvation shells is required, and the dependence of the results on the number of solvation shells included in the exchange calculation. For the larger Ala_{10}, the computational demands of obtaining high-precision data for various hybrid models (which require fully solvated simulations) prevented exhaustive testing. Thus, these more detailed tests were performed on the smaller models alanine dipeptide (blocked Ala_1) and alanine tetrapeptide (blocked Ala_3).

Alanine Dipeptide. We first compared results obtained for the standard REMD with TIP3P to those from 2 different GB models as well as to TIP3P but using the hybrid solvent model for calculation of exchange probability. The hybrid model employed either a first solvent shell (30 TIP3P waters) or first and second shells (60 waters). The population of minima corresponding to alternate secondary structure types (see Methods for details) are shown in Table 1. The largest population is found for the polyproline II basin (~35%), followed by an α-helix and a β-sheet (each ~25%), and a much lower population of a left-handed α-helix or turn conformation (1−3%). We make the observation that all of these solvent models provide essentially the same results. Use of either GB^{OBC} or GB^{HCT} with no explicit solvent either in MD or in the exchange calculation provides populations for each of the basins with an error of ~2% population as compared to the standard REMD in the explicit solvent. Similarly, the average SASA is nearly identical for all models. These data indicate that the hybrid model is at least performing adequately and does not have any obvious and serious problems and that similar results are obtained for either the first and second solvation shells or only the first shell. This insensitivity is expected since the GB simulations adequately reproduced the explicit solvent data with no explicit solvent shell. The insensitivity of the results to solvent model strongly indicates that alanine dipeptide is not a good test case for evaluation of the effects of inclusion of explicit solvent.

Alanine Tetrapeptide. We next turn to results from alanine tetrapeptide to evaluate whether the agreement

Table 1. Populations of Basins on the Alanine Dipeptide ϕ/ψ Energy Landscape Corresponding to Alternate Secondary Structures, along with Average Solvent Accessible Surface Areas[a]

alanine dipeptide	α	β	P^{II}	α^L	SASA
explicit solvent	28.1 ± 1.0	25.1 ± 0.1	36.2 ± 0.5	2.6 ± 0.1	355.8 ± 0.0
GB^{OBC}	29.3 ± 0.8	26.5 ± 0.5	35.1 ± 0.2	0.7 ± 0.1	356.5 ± 0.0
GB^{HCT}	28.5 ± 0.2	27.6 ± 0.1	34.0 ± 0.2	0.8 ± 0.2	356.5 ± 0.1
hybrid first shell + GB^{OBC}	29.7 ± 1.8	24.7 ± 0.4	35.0 ± 1.5	2.5 ± 0.1	355.8 ± 0.1
hybrid first and second shells + GB^{OBC}	30.3 ± 1.5	24.7 ± 0.3	36.0 ± 0.2	1.3 ± 0.8	355.9 ± 0.1

[a] The results for the pure GB and hybrid REMD models are all similar to those obtained using standard REMD with full explicit solvent.

Table 2. Data for the Central Alanine in Alanine Tetrapeptide (Blocked Ala_3)[a]

alanine tetrapeptide	α	β	P^{II}	α^L	SASA
explicit solvent	23.6 ± 0.1	23.4 ± 1.3	40.2 ± 1.4	5.1 ± 0.1	565.3 ± 0.1
GB^{OBC}	50.5 ± 2.4	17.5 ± 0.9	22.9 ± 0.6	1.1 ± 0.4	557.4 ± 1.0
GB^{HCT}	57.8 ± 1.0	15.2 ± 0.2	18.2 ± 0.4	1.2 ± 0.1	552.4 ± 0.4
hybrid first shell noGB	41.4 ± 0.8	13.5 ± 0.9	23.4 ± 1.0	13.1 ± 0.8	552.7 ± 0.1
hybrid first and second shells noGB	29.5 ± 0.2	14.1 ± 0.2	24.1 ± 0.5	23.4 ± 0.3	550.8 ± 0.2
hybrid first shell GB^{OBC}	21.6 ± 0.9	21.2 ± 0.2	41.1 ± 0.3	7.6 ± 1.0	563.2 ± 0.1
hybrid first and second shells GB^{OBC}	28.3 ± 1.7	22.2 ± 0.9	37.7 ± 0.2	3.8 ± 0.1	563.8 ± 0.2
hybrid first shell + GB^{HCT}	23.5 ± 1.1	22.1 ± 0.8	42.8 ± 1.0	2.3 ± 0.0	566.4 ± 0.2
hybrid first and second shells + GB^{HCT}	14.9 ± 0.2	25.6 ± 0.1	49.4 ± 0.4	1.9 ± 0.4	569.6 ± 0.1

[a] Populations of basins on the ϕ/ψ energy landscape corresponding to alternate secondary structures are shown, along with average solvent accessible surface areas. Data are discussed in the text.

between all solvent models tested for alanine dipeptide is maintained in larger systems. In Table 2 we show populations for secondary structure basins for the central alanine residue using standard REMD with explicit solvent, GB^{OBC} or GB^{HCT}. Data are also shown for several hybrid models, as discussed below.

For standard REMD in explicit solvent, we observe that the populations have not changed significantly from those obtained for alanine dipeptide, with a slight increase in population of the polyproline II conformation that dominates the ensemble. In this case, however, we observe that both of the pure GB models are in significant disagreement with TIP3P, with α-helical conformations dominating the ensemble (over 50% for each GB model). The two GB models are similar to each other. Overstabilization of salt bridges in GB has been reported,[12,26,27] but no salt bridges are present in this system.

Next, we performed REMD simulations in explicit solvent, but retain only the first (50) or the first and second (100) solvation shells in the exchange calculation. Importantly, no GB model was included in these simulations. Using only a single solvation shell results in a significant bias in favor of α-helical conformations (41% vs ∼24% for standard REMD), much too little polyproline II conformation and nearly three times the α^L/turn conformation than was sampled in standard REMD. Inclusion of a second shell (without GB) resulted in an even greater shift of the ensemble toward turn structures. Notably, both of these shell models show a significantly smaller average SASA than obtained with standard REMD in the explicit solvent, consistent with a drive toward compact conformations that reduce the water/vacuum interface that is present without a reaction field to surround the solvent shells.

We next examine the data obtained from the hybrid model in which GB solvation was employed in addition to shells of explicit solvation. We note that all of these models are in significantly better agreement with the standard TIP3P REMD data, regardless of the GB method or number of shells. The more recent GB^{OBC} model performed best, with errors in population of only ∼3% for all basins with the exception of the α-helix conformation with the first and second shell model, which had an error that was less than 5%. The average SASA was also in excellent agreement with standard REMD. We conclude that this hybrid model is significantly better than the pure GB REMD or inclusion of only the solvation shells with no reaction field. The addition of a second shell in the exchange calculation appears to make no significant difference as compared to a single shell.

As described above, the MD simulations between exchanges in the hybrid model are performed with full explicit solvation. We thus do not need to restrain the explicit water, and since the solvation shells are surrounded by bulk explicit solvent, we expect no effect on the water geometries as have been reported when using a hybrid GB+explicit water model for dynamics.[41] To test this hypothesis, we calculated the radial distribution function for water oxygens around the carbonyl oxygen in the central Ala2 and found that the function obtained in the hybrid model was indistinguishable from that in the standard REMD in the explicit solvent (Figure S1). Since these data are obtained from the entire set of structures, this close agreement is also a further indicator of the similarity of the ensembles obtained using hybrid or standard REMD.

The hybrid model using GB^{HCT} performed comparably to GB^{OBC} when only a single shell was used, but the first+second shell model showed a marked reduction in the α-helix conformation (from 23.5% to 14.9%). This was accompanied by an increase in the average SASA. These effects with GB^{HCT} are even more apparent in Ala_{10} and will be discussed in more detail below.

Polyalanine (Ala_{10}). The conformational variability available to Ala_{10} is significantly greater than for alanine dipeptide

Table 3. Data for the Central Ala5 in Blocked Ala_{10}[a]

Ala_{10}	α	β	P^{II}	α^L	SASA
explicit solvent	24.9 ± 0.8	19.5 ± 0.6	39.5 ± 0.4	8.4 ± 2.0	1195.4 ± 5.6
GB^{OBC}	67.8 ± 1.8	8.3 ± 0.7	12.5 ± 0.8	4.2 ± 0.1	1098.6 ± 0.4
GB^{HCT}	83.1 ± 0.1	3.2 ± 0.1	5.0 ± 0.0	2.3 ± 0.1	1038.3 ± 1.6
hybrid GB^{OBC} +first shell	35.7 ± 6.2	17.3 ± 0.2	29.0 ± 5.3	6.6 ± 0.7	1140.8 ± 4.4
hybrid GB^{HCT} + first shell	12.3 ± 0.2	28.3 ± 0.3	50.5 ± 1.2	2.1 ± 1.1	1275.4 ± 2.5
hybrid $GB^{OBC'}$ + first shell	29.8 ± 1.6	18.5 ± 1.6	34.3 ± 0.5	8.9 ± 0.3	1167.8 ± 2.5

[a] Populations of basins on the ϕ/ψ energy landscape corresponding to alternate secondary structures are shown, along with average solvent accessible surface areas. $GB^{OBC'}$ refers to the hybrid model using GB^{OBC} with slight adjustment of the Born radius on H bonded to O. Uncertainties reflect differences between independent simulations from different initial structures. Data are discussed in the text.

or tetrapeptide. We thus performed a more stringent evaluation of data convergence in this case to ensure that the differences we observe between the different solvent models are statistically significant. We performed two completely independent REMD simulations for each of the solvent models, in each case starting from 2 different initial ensembles (fully extended or fully helical). This allows us to evaluate the influence of the solvent model within the context of intrinsic uncertainties in each data set.

We also consider separately the local ϕ/ψ conformations and more global properties of this larger peptide, such as end-to-end distance distributions and conformation cluster analysis.

Comparison of Local Conformational Preferences. In Table 3 we show secondary structure basin populations for the central Ala5 residue. Free energy surfaces for these simulations are provided in Figure S2. For the reference standard REMD simulations in explicit solvent, the polyproline II conformation is again favored with the same ∼40% population as we obtained for alanine dipeptide and tetrapeptide. In comparison, both GB models show a very large bias in favor of α-helix conformations (∼70−80%).

Consistent with the results obtained for alanine tetrapeptide, the GB^{HCT} hybrid model favors extended conformations with large SASA too strongly (β and P_{II}), despite the bias in favor of an α-helix for the pure GB^{HCT} simulations. This suggests that the explicit water shell is solvated too strongly by this GB model. The GB^{OBC} hybrid model shows a more balanced profile in good agreement with the full TIP3P data. The strong bias favoring an α-helix in the pure GB^{OBC} model is nearly completely eliminated when a single solvent shell is retained, although some remains with approximately 10% too much α-helix present in the GB^{OBC} hybrid.

In addition to differences in the method for calculating GB effective Born radii, the GB^{HCT} and GB^{OBC} simulations employed different intrinsic Born radii (denoted in Amber as mbondi and mbondi2 sets, respectively), consistent with recommendations for these models. To determine the relative influence of these two differences, we repeated the calculations, swapping the GB models and radii (GB^{HCT} with mbondi2, GB^{OBC} with mbondi). We found that the results depended nearly exclusively on the set of radii and were less sensitive to the GB models themselves (data not shown). This is consistent with the aim of the GB^{OBC} model, which was designed to provide improved properties for larger systems than our current model.[58] We note that the strong bias toward extended structures seen in the hybrid models using mbondi radii likely arises from the use of 0.8 Å for hydrogen atoms bonded to oxygen. In the more recent mbondi2 set, this value was restored to the default Bondi value of 1.2 Å. This larger value appears to have an improved balance of hydrogen bonding of the explicit solvent to the solute or to the bulk (continuum) solvent.

Comparison of Global Structural Properties. Our analysis of alanine dipeptide and tetrapeptide focused on local backbone conformation; in the larger Ala_{10} we supplement this analysis with more global properties of the chain. We calculated the end-to-end distance distributions for Ala_{10} in the 300 K ensembles obtained from each of the different REMD simulations. In Figure 6 we show the results of the 2 explicit solvent REMD simulations that were initiated from fully α-helical or extended conformations, respectively. A broad distribution of distances is observed, suggesting that no particular conformation is preferred, consistent with the local backbone preferences for the central Ala5. Consistent with the small uncertainties in the ϕ/ψ basin populations, we observe that the initial conformation has essentially no effect on the distribution, indicating that the REMD simulations are well-converged on this time scale. Similar behavior is observed for other temperatures. As expected, standard MD simulations at 300 K were trapped near the initial conformation on this time scale (data not shown).

In Figure 6, we show the distance distributions at 300 K obtained from GB REMD using the two GB models (HCT and OBC). In contrast to the relatively flat profiles seen in the explicit solvent REMD data, a sharp peak near 11 Å is obtained using either GB model, with essentially no sampling of extended conformations with end-to-end distances greater than ∼15−20 Å, unlike the explicit solvent REMD that shows a nearly flat distribution out to ∼22 Å. This is consistent with the strong bias toward α-helix in the pure GB models as shown in Table 3. The bias is somewhat less pronounced with the GB^{OBC} model than with GB^{HCT}. We note that these differences between the various solvent models are much larger than the differences obtained from alternate initial conformations using the same solvent model.

In Figure 6 we also show end-to-end distance distributions at 300 K obtained from REMD with the same hybrid variations shown in Table 3, each of which retained only the first shell (100 closest) water molecules combined with different GB models in the exchange calculation. When GB^{HCT} was used in the hybrid model (Figure 6C), the distributions differ significantly from the reference explicit solvent REMD data, consistent with the large increase in polyproline II backbone conformations and average SASA

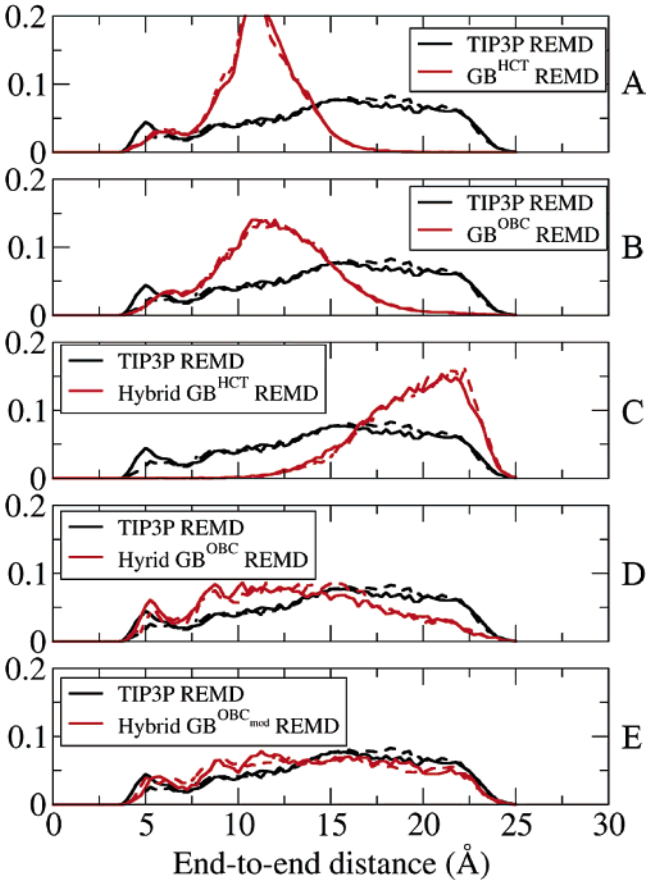


Figure 6. Ala_{10} end-to-end distance distributions at 300 K obtained in REMD using alternate solvent models (red): (A) pure GB^{HCT}, (B) pure GB^{OBC}, (C) hybrid REMD with GB^{HCT} and mbondi radii, (D) hybrid REMD with GB^{OBC} and mbondi2 radii ($H^O = 1.2$ Å), and (E) hybrid REMD with $GB^{OBC'}$ (mbondi2 radii with $H^O = 1.15$ Å). In each case the results are independent of initial conformation (solid/dashed lines). Data from standard REMD with explicit solvent are shown in each graph for comparison (black).

for this model shown in Table 3. This bias toward more extended conformations in the hybrid using GB^{HCT} is also consistent with what we observed for alanine tetrapeptide (Table 2).

We next analyzed the distributions obtained from the GB^{OBC} hybrid model (Figure 6D). In this case, much better agreement with the reference data is seen than with either GB^{OBC} alone or the explicit/GB^{HCT} hybrid. However, the sampling of the most extended conformations (longest end-to-end distances) is slightly reduced in the hybrid REMD simulations.

The good convergence of our data suggested the possibility of using it for minor empirical adjustment of the mbondi2 values for use with the GB^{OBC} hybrid model. We adjusted the radii of hydrogen bonded to either N or O by 0.05 Å. Modification of H on N had little effect on the resulting distributions (data not shown), but reduction of the radius of H on O from 1.2 to 1.15 Å ($GB^{OBC'}$) resulted in an end-to-end distance distribution in improved agreement with standard explicit solvent REMD data (Figure 6E and Table 3). This slight reduction in the hydrogen radius is consistent with the increased electronegativity of oxygen.[71] This change

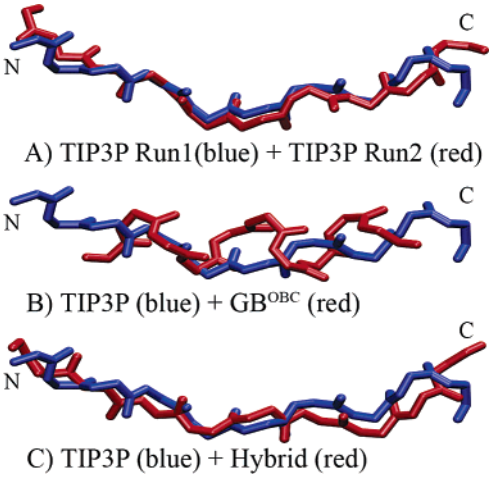


Figure 7. Representative structures for the most populated clusters in 300 K ensembles obtained using various solvent models. (A) Very similar P_{II} structures are obtained from 2 independent standard REMD simulations with explicit solvent, initiated in extended and fully helical conformations. (B) Comparison of structures from GB^{OBC} and TIP3P. GB^{OBC} prefers α-helical conformations, in disagreement with explicit solvent simulations. (C) Using $GB^{OBC'}$ with the hybrid model provides structures in close agreement with standard REMD in TIP3P. Terminal residues were not included in the cluster analysis.

does not affect the pure GB calculations since Ala_{10} has no H bonded to O.

The $GB^{OBC'}$ hybrid model showed improved agreement with the pure TIP3P data, with all basin populations within 5% of the standard explicit solvent REMD. Some slight bias favoring an α-helix at the expense of some polyproline II conformation remains in this model and will be the subject of future investigation. We repeated the simulations of alanine dipeptide and tetrapeptide using this modified radius and found that the populations (Table S2) remained in good agreement with standard REMD with explicit solvent.

Since the backbone conformation populations suggest that the P_{II} basin is the global free energy minimum in both the standard explicit solvent and the hybrid solvent models (Table 3 and Figure S2), we performed cluster analysis to determine the extent to which this local preference was reflected in the conformation of the entire polymer chain. Once again we compare results from independent ensembles generated by REMD with different initial conformations to ensure the convergence of our data.

The most populated cluster for Ala_{10} at 300 K in both standard explicit solvent REMD runs was an extended P_{II} conformation (over 98% of the local backbone conformations in this cluster are P_{II}, data not shown). This fully P_{II} cluster comprised ~20% of the overall ensemble in both explicit solvent simulations (19.5% vs 21.2%). Representative structures for the clusters obtained from the independent simulations differed by only 1.3 Å in backbone RMSD (Figure 7A). Once again, the high level of consistency between the data sets and independence of not only the conformation but also the absolute population of the clusters give us confidence in the converged nature of our data. The relatively low population of this cluster in both simulations is also consistent

with the broad distribution of end-to-end distances (Figure 6). A more detailed analysis of the ensemble of structures sampled by Ala_{10} will be presented elsewhere, but this preference for P_{II} conformations is consistent with the experimental and simulation reports described previously.

As was demonstrated with the analyses presented above, the pure GB^{HCT} and GB^{OBC} REMD simulations do not reproduce the data obtained in the explicit solvent, nor are they consistent with experimental data. The most populated cluster in both cases is fully α-helical (Figure 7B shows the GB^{OBC} structure), comprising ∼48% of the overall ensemble for GB^{HCT} and 25.4% for GB^{OBC}. This analysis is consistent with the α-helical bias apparent in the Ramachandran free energy surfaces shown in Figure S2.

We next performed cluster analysis on the ensembles obtained with the $GB^{OBC'}$ hybrid model with modified mbondi2 radii. Consistent with the standard explicit solvent REMD runs, the most populated cluster at 300 K was also an extended P_{II} conformation. Representative structures were within 1.5 Å backbone RMSD from those obtained in the explicit solvent (Figure 7C), again suggesting that the hybrid model is able to capture the dominant effects of the explicit solvent in the exchange calculation despite the need for many fewer replicas.

Since the most populated clusters were in close agreement between both TIP3P REMD simulations and the $GB^{OBC'}$ hybrid model, we compared the populations of all clusters observed. Smith et al. showed[73] that cluster analysis of simulations was a much more stringent test of convergence than other measures that they tested, including energy, RMSD, or diversity of hydrogen bonds sampled. This was particularly useful when analyzing coordinate sets obtained by merging two independent trajectories. They examined the 5 ns dynamics of an 11-residue peptide and showed that the two trajectories sampled essentially none of the same clusters.

We adapted this approach to our analysis, but we emphasize not only just the existence of conformation families in two data sets but also the fractional population of each cluster in 300 K ensembles sampled in independent simulations. All trajectories from TIP3P REMD, GB^{OBC} REMD, and hybrid $GB^{OBC'}$ simulations were combined, and the resulting data set was clustered. A total of 44 clusters contained 99% of the structures; the fraction of the ensemble corresponding to each cluster was calculated for each REMD simulation. We compared the population of each cluster in the different ensembles, including those generated with the same or different solvent models.

First we evaluated the convergence of our standard REMD simulations with TIP3P by comparing cluster sizes between the independent runs with different initial conformations (extended and fully α-helical). Not only were the same conformations sampled in each run (20.3±0.9%), but the populations of clusters in each ensemble were highly correlated (Figure 8A, R^2=0.974 and a slope of 1.02). This indicates that the relative population of each structure type is highly converged in these data sets.

In stark contrast, when the TIP3P and GB^{OBC} ensembles are compared, no correlation between cluster populations is observed (Figure 8B, R^2=0.075), and the largest cluster in each (∼20%) has less than 2% population in the other model. Much better results are obtained from the $GB^{OBC'}$ hybrid data, with a correlation coefficient of 0.935 with the standard TIP3P REMD data (Figure 8C). All clusters larger than 5% have the same rank order in the two models. There is a relatively small difference in the size of the single cluster that is the largest for both models (15.9±0.6% and 20.3±0.9% for hybrid and standard TIP3P REMD, respectively). This corresponds to an error of only 0.15 kcal/mol for the free energy of this cluster between the two models, compared to the 0.05 kcal/mol difference obtained between data sets from the same model. For comparison, the error in the free energy of this conformation using GB was more than 10 times larger (1.6 kcal/mol).

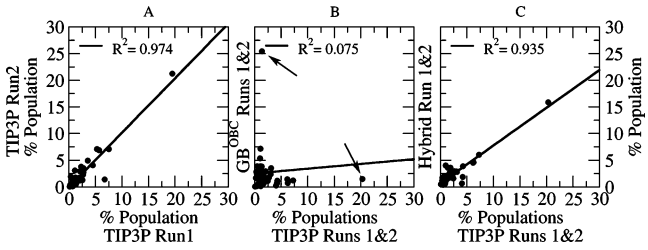


Figure 8. Cluster populations at 300 K from REMD for TIP3P Run1 vs Run2 (A), TIP3P Runs 1&2 vs GB^{OBC} Runs 1&2 (B), and TIP3P Runs 1&2 vs hybrid $GB^{OBC'}$ Runs 1&2. High correlations between individual TIP3P simulations and between TIP3P and hybrid simulations are observed, with the difference in the largest cluster in (C) corresponding to an error in free energy of only 0.15 kcal/mol. No correlation between TIP3P and GB^{OBC} is observed; note also in plot (B) that the largest cluster in each solvent model has very low population in the other model (indicated by arrows).

Since the standard explicit solvent REMD and hybrid solvent using $GB^{OBC'}$ have the same most populated cluster, we investigated the time scale required for each model to adopt this conformation as the dominant member of their ensemble. This is important since the standard REMD simulation employed many more replicas, possibly facilitating an earlier location of the P_{II} conformation that would then be adopted in the lowest temperature ensembles. In Figure 9 we show the fractional size of this cluster in the structures sampled as a function of time for the standard REMD and the hybrid REMD, including data from both initial conformations in each model. Data are shown at 300 K, and the first 5 ns were discarded in each case to remove biasing of the populations by the initial conformations that were not sampled at later points. The level of agreement is impressive; the long-time averages for both simulations of the 2 models are all ∼20%, with convergence to this value occurring at approximately 5 ns in all cases (in addition to the 5 ns that were discarded).

Conclusions

We introduced a new variant of replica exchange molecular dynamics in which simulations are performed with a fully explicit representation of the solvent, but those solvent molecules beyond the first solvation shell are replaced with a continuum description only for the purpose of calculating the exchange probability. This reduces the effective system

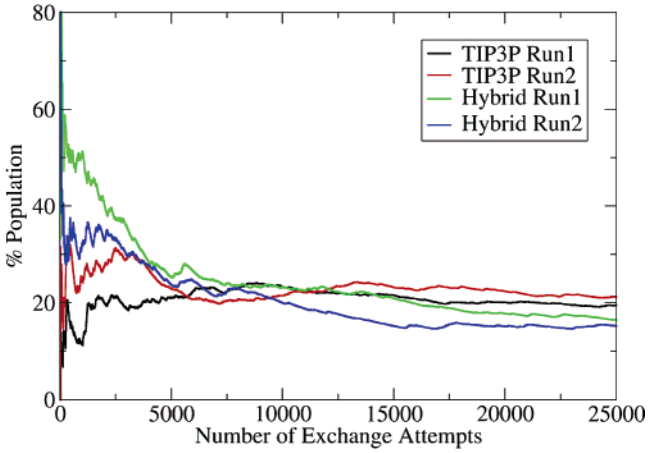


Figure 9. Population of the cluster corresponding to polyproline II helix (Figure 7) as a function of time for REMD simulations in explicit solvent, with the 2 independent simulations using the full system energy in the exchange calculation shown in black/red and the GB^{OBC} hybrid shown in green/blue. At ~5 ns, all four simulations converge to a population of 16−20% (the largest cluster in each of the ensembles), with a slightly lower population in the hybrid models that is consistent with Figure 8C.

size governing the number of replicas required to span a given temperature range and therefore significantly reduces the computational cost of REMD simulations. This approach is similar in spirit to hybrid explicit/continuum models that have been proposed for use during each step of MD simulation; in the present case, however, the solvent is fully explicit during the dynamics, and no restraints are needed to maintain a solvation shell. However, since the Hamiltonian used for the exchange differs from that employed during dynamics, these simulations are approximate and are not guaranteed to provide correct canonical ensembles. It is important to determine the extent to which this approximation affects the resulting ensembles; in this article we introduce the method and investigate some of these effects on several short alanine-based peptides.

Recently, another approach to reducing the number of replicas required for explicit solvent REMD simulations was proposed[74] in which the water−water interaction energy was temperature-dependent. That study employed alanine dipeptide as a model to show that their less computationally demanding method provided a similar ensemble to that obtained with the standard REMD. In the present work we show that alanine dipeptide conformations are nearly insensitive to the solvent models that we tested, with results from the full explicit solvent, two different GB models, and several hybrid models all providing similar ensembles. In contrast, several of these models provided ensembles for the longer peptides that were in significant disagreement with the standard REMD in the explicit solvent, indicating that larger model systems should be included in evaluation of solvent models.

We further tested the method by calculation of conformational ensembles of Ala_{10} using the TIP3P explicit solvent model, two GB models available in Amber, and hybrid variants using TIP3P and each GB model, all using the same underlying protein force field parameters. Ensembles from standard REMD in the explicit solvent were considered the standard, and convergence of this data set was validated by a high correlation (R^2=0.974) between the fractional populations of conformation families in simulations initiated with completely different initial structure ensembles. While a broad distribution of conformations was sampled, the predominant cluster for Ala_{10} adopted a P_{II} structure. This preference is consistent with reported experimental and computational results for short polyalanine peptides.[75]

Simulations using the hybrid model with GB^{OBC} were in excellent agreement with the reference data for local backbone conformations, end-to-end distance, SASA, and populations of each conformation family in the ensemble. The difference in population in the largest cluster indicates that the hybrid model introduced an error of less than 0.2 kcal/mol in free energy while reducing the computational expense by a factor of 5.

In contrast, REMD using only the GB models provided ensembles that bore no resemblance to the reference data, with the GB ensembles incorrectly dominated by α-helical conformations. This may be indicative of general errors in these GB models, or they may arise from neglect of the structure in the first solvation shells of the peptide. Mezei et al. recently reported[55] free energy calculations using explicit solvent, showing that solvation strongly favors the P_{II} conformation over an α-helix. Solvation free energy was shown to be highly correlated with the energy of interaction between the peptide and its first solvation shell.

It is important to note that several challenges remain for more general use of the proposed hybrid approach. In particular, the present work studied the effects on alanine-based peptides. Future studies should be performed on other sequences with a more diverse representation of functional groups in the side chains. In particular, it will be important to determine whether the hybrid model is able to overcome known issues with GB models and ions pair interactions. The inclusion of explicit counterions in the exchange calculation may also be problematic. Additionally, we demonstrated that inclusion of a single shell of explicit water was sufficient for alanine dipeptide and analine tetrapeptide. In both cases similar results were obtained using one or two shells, but we were unable to perform these comparisons for Ala_{10}. Although our approach reduces the number of replicas required for REMD, the simulations are still fully solvated during each step of MD and obtaining well converged data requires a significant investment of computational resources.

The results obtained from these model systems provide additional evidence that explicit representation of water in the first solvation shell can significantly improve the performance of the GB continuum models, providing data similar to standard REMD with a fully explicit solvent but at a greatly reduced cost. This reduction in computational requirements can enable simulations on longer time scales for the same system size or permit application of REMD to the study of much larger systems. We also showed that use of one or two explicit solvent shells alone was inadequate and that adding a reaction field was essential for obtaining

reasonable results. Adaptation of this method to other continuum models (such as the more rigorous PB) should be straightforward. Since the continuum solvent is only used for the infrequent exchange calculations, models that are too complex for use at each step of dynamics can be readily employed.

Acknowledgment. The authors thank Adrian Roitberg for helpful feedback, John Mongan and Alexey Onufriev for valuable discussions concerning the GB models, and Guanglei Cui for help with Amber REMD. Roberto Gomperts provided important code optimizations for Amber. Supercomputer time at NCSA (NPACI MCA02N028) and financial support from the National Institutes of Health (NIH GM6167803) and Department of Energy (Contract DE-AC02-98CH10886) are gratefully acknowledged. Additional computer time was generously provided by the SGI Engineering group. C.S. is a Cottrell Scholar of Research Corporation.

Supporting Information Available: Basin populations for Ala_1 and Ala_3 using hybrid $GB^{OBC'}$, free energy profiles for backbone conformations in Ala5 for Ala_{10}, radial distribution functions for water near Ala_3, and ranges for definition of secondary structure basins. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Tai, K. *Biophys. Chem.* **2004**, *107* (3), 213−220.

(2) Roitberg, A.; Simmerling, C. *J. Mol. Graphics Modell.* **2004**, *22* (5), 317−317.

(3) Hansmann, U. H. E. *Chem. Phys. Lett.* **1997**, *281* (1−3), 140−150.

(4) Swendsen, R. H.; Wang, J. S. *Phys. Rev. Lett.* **1986**, *57* (21), 2607−2609.

(5) Tesi, M. C.; vanRensburg, E. J. J.; Orlandini, E.; Whittington, S. G. *J. Stat. Phys.* **1996**, *82* (1−2), 155−181.

(6) Sugita, Y.; Okamoto, Y. *Chem. Phys. Lett.* **1999**, *314* (1−2), 141−151.

(7) Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. *J. Chem. Phys.* **1953**, *21*, 1087−1092.

(8) Feig, M.; Karanicolas, J.; Brooks, C. L. *J. Mol. Graphics Modell.* **2004**, *22* (5), 377−395.

(9) Garcia, A. E.; Sanbonmatsu, K. Y. *Proteins: Struct., Funct., Genet.* **2001**, *42* (3), 345−354.

(10) Garcia, A. E.; Sanbonmatsu, K. Y. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99* (5), 2782−2787.

(11) Karanicolas, J.; Brooks, C. L. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100* (7), 3954−3959.

(12) Pitera, J. W.; Swope, W. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100* (13), 7587−7592.

(13) Sugita, Y.; Kitao, A.; Okamoto, Y. *J. Chem. Phys.* **2000**, *113* (15), 6042−6051.

(14) Zhou, R.; Berne, B. J.; Germain, R. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98* (26), 14931−6.

(15) Kinnear, B. S.; Jarrold, M. F.; Hansmann, U. H. E. *J. Mol. Graphics Modell.* **2004**, *22* (5), 397−403.

(16) Roe, D. R.; Hornak, V.; Simmerling, C. *J. Mol. Biol.* **2005**, *352* (2), 370−381.

(17) Rathore, N.; Chopra, M.; de Pablo, J. J. *J. Chem. Phys.* **2005**, *122* (2), 024111.

(18) Fukunishi, H, W. O., Takada, S. *J. Chem. Phys.* **2002**, *116* (20), 9058−9067.

(19) Cheng, X. L.; Cui, G. L.; Hornak, V.; Simmerling, C. *J. Phys. Chem. B* **2005**, *109* (16), 8220−8230.

(20) Kofke, D. A. *J. Chem. Phys.* **2002**, *117* (15), 6911−6914.

(21) Sugita, Y.; Okamoto, Y. *Chem. Phys. Lett.* **2000**, *329* (3−4), 261−270.

(22) Mitsutake, A., S. Y., Okamoto, Y. *J. Chem. Phys.* **2003**, *118* (14), 6664−6688.

(23) Jang, S.; Shin, S.; Pak, Y. *Phys. Rev. Lett.* **2003**, *91* (5), 58305.

(24) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. *J. Am. Chem. Soc.* **1990**, *112* (16), 6127−6129.

(25) Nymeyer, H.; Garcia, A. E. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100* (24), 13934−13939.

(26) Zhou, R.; Berne, B. J. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99* (20), 12777−82.

(27) Zhou, R. *Proteins* **2003**, *53* (2), 148−61.

(28) Simmerling, C.; Strockbine, B.; Roitberg, A. *J. Am. Chem. Soc.* **2002**, *124* (38), 11258.

(29) Ghosh, A.; Rapp, C. S.; Friesner, R. A. *J. Phys. Chem. B* **1998**, *102* (52), 10983−10990.

(30) Srinivasan, J.; Cheatham, T. E.; Cieplak, P.; Kollman, P. A.; Case, D. A. *J. Am. Chem. Soc.* **1998**, *120* (37), 9401−9409.

(31) Luo, R.; David, L.; Gilson, M. K. *J. Comput. Chem.* **2002**, *23* (13), 1244−1253.

(32) Jeancharles, A.; Nicholls, A.; Sharp, K.; Honig, B.; Tempczyk, A.; Hendrickson, T. F.; Still, W. C. *J. Am. Chem. Soc.* **1991**, *113* (4), 1454−1455.

(33) Alper, H.; Levy, R. M. *J. Chem. Phys.* **1993**, *99* (12), 9847−9852.

(34) Beglov, D.; Roux, B. *Biopolymers* **1995**, *35* (2), 171−178.

(35) Beglov, D.; Roux, B. *J. Chem. Phys.* **1994**, *100* (12), 9050−9063.

(36) Brooks, C. L.; Brunger, A.; Karplus, M. *Biopolymers* **1985**, *24* (5), 843−865.

(37) Brooks, C. L.; Karplus, M. *J. Chem. Phys.* **1983**, *79* (12), 6312−6325.

(38) Kentsis, A.; Mezei, M.; Gindin, T.; Osman, R. *Proteins: Struct., Funct., Bioinformatics* **2004**, *55* (3), 493−501.

(39) King, G.; Warshel, A. *J. Chem. Phys.* **1989**, *91* (6), 3647−3661.

(40) Lee, M. S.; Olson, M. A. *J. Phys. Chem. B* **2005**, *109* (11), 5223−5236.

(41) Lee, M. S.; Salsbury, F. R.; Olson, M. A. *J. Comput. Chem.* **2004**, *25* (16), 1967−1978.

(42) Das, B.; Helms, V.; Lounnas, V.; Wade, R. C. *J. Inorg. Biochem.* **2000**, *81* (3), 121−131.

(43) Topol, I. A.; Tawa, G. J.; Burt, S. K.; Rashin, A. A. *J. Chem. Phys.* **1999**, *111* (24), 10998−11014.

(44) van der Spoel, D.; van Maaren, P. J.; Berendsen, H. J. C. *J. Chem. Phys.* **1998**, *108* (24), 10220−10230.

(45) Vorobjev, Y. N.; Hermans, J. *Biophys. Chem.* **1999**, *78* (1−2), 195−205.

(46) Errington, N.; Doig, A. J. *Biochemistry* **2005**, *44* (20), 7553−8.

(47) Groebke, K.; Renold, P.; Tsang, K. Y.; Allen, T. J.; McClure, K. F.; Kemp, D. S. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93* (9), 4025−9.

(48) Marqusee, S.; Robbins, V. H.; Baldwin, R. L. *Proc. Natl. Acad. Sci. U.S.A.* **1989**, *86* (14), 5286−5290.

(49) Maison, W.; Arce, E.; Renold, P.; Kennedy, R. J.; Kemp, D. S. *J. Am. Chem. Soc.* **2001**, *123* (42), 10245−54.

(50) Heitmann, B.; Job, G. E.; Kennedy, R. J.; Walker, S. M.; Kemp, D. S. *J. Am. Chem. Soc.* **2005**, *127* (6), 1690−704.

(51) Chen, K.; Liu, Z. G.; Kallenbach, N. R. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101* (43), 15352−15357.

(52) McColl, I. H.; Blanch, E. W.; Hecht, L.; Kallenbach, N. R.; Barron, L. D. *J. Am. Chem. Soc.* **2004**, *126* (16), 5076−5077.

(53) Shi, Z. S.; Olson, C. A.; Rose, G. D.; Baldwin, R. L.; Kallenbach, N. R. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99* (14), 9190−9195.

(54) Asher, S. A.; Mikhonin, A. V.; Bykov, S. *J. Am. Chem. Soc.* **2004**, *126* (27), 8433−8440.

(55) Mezei, M.; Fleming, P. J.; Srinivasan, R.; Rose, G. D. *Proteins: Struct., Funct., Bioinformatics* **2004**, *55* (3), 502−507.

(56) Garcia, A. E. *Polymer* **2004**, *45* (2), 669−676.

(57) Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. *Chem. Phys. Lett.* **1995**, *246* (1−2), 122−129.

(58) Feig, M.; Onufriev, A.; Lee, M. S.; Im, W.; Case, D. A.; Brooks, C. L. *J. Comput. Chem.* **2004**, *25* (2), 265−284.

(59) Onufriev, A.; Bashford, D.; Case, D. A. *J. Phys. Chem. B* **2000**, *104* (15), 3712−3720.

(60) Kofke, D. A. *J. Chem. Phys.* **2004**, *121* (2), 1167−1167.

(61) 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* (19), 5179−5197.

(62) Wang, J. M.; Cieplak, P.; Kollman, P. A. *J. Comput. Chem.* **2000**, *21* (12), 1049−1074.

(63) Hornak, V.; Simmerling, C. Manuscript in preparation.

(64) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926−935.

(65) Case, D. A.; Cheatham, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. *J. Comput. Chem.* **2005**, *26* (16), 1668−1688.

(66) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *23* (3), 327−341.

(67) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Dinola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81* (8), 3684−3690.

(68) Simmerling, C.; Elber, R.; Zhang, J., MOIL-View − A Program for Visualization of Structure and Dynamics of Biomolecules and STO − A Program for Computing Stochastic Paths. In *Modelling of Biomolecular Structures and Mechanisms*; Pullman et al., A., Ed.; Kluwer Academic Publishers: Netherlands, 1995; pp 241−265.

(69) Darden, T.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993**, *98* (12), 10089−10092.

(70) Bondi, A. *J. Phys. Chem.* **1964**, *68* (3), 441−451.

(71) Tsui, V.; Case, D. A. *J. Am. Chem. Soc.* **2000**, *122* (11), 2489−2498.

(72) Ponder, J. W.; Richards, F. M. *J. Comput. Chem.* **1987**, *8* (7), 1016−1024.

(73) Smith, L. J.; Daura, X.; van Gunsteren, W. F. *Proteins: Struct., Funct., Genet.* **2002**, *48* (3), 487−496.

(74) Liu, P.; Kim, B.; Friesner, R. A.; Berne, B. J. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102* (39), 13749−13754.

(75) Shi, Z.; Woody, R. W.; Kallenbach, N. R. *Adv. Prot. Chem.* **2002**, *62*, 163−240.

CT050196Z

JCTC Journal of Chemical Theory and Computation

Toward a Theoretical Quantitative Estimation of the λ_{max} of Anthraquinones-Based Dyes

Eric A. Perpète,*[,†,‡] Valerie Wathelet,[‡] Julien Preat,[‡] Christophe Lambert,[§] and Denis Jacquemin*[,†,‡]

Laboratoire de Chimie Théorique Appliquée, Facultés Universitaires Notre-Dame de la Paix, rue de Bruxelles, 61, B-5000 Namur, Belgium, and BioXpr, Centre technologique FUNDP, Rue du séminaire, 22, B-5000 Namur, Belgium

Received November 23, 2005

Abstract: We have computed the absorption spectra of a large series of anthraquinone dyes by using the time-dependent density functional theory (TD-DFT) for the excited-state calculations and the polarizable continuum model (PCM) for evaluating bulk solvent effects. On one hand, we compare the results obtained with the B3LYP and the PBE0 hybrid functionals, combined with different atomic basis sets. On the other hand, using multiple linear regression, we take advantage of the λ_{max} predicted by these two functionals in order to reach the best agreement between theoretical estimates and experimental measurements. It turns out that 1. PBE0 provides more accurate results than B3LYP; in addition the average errors provided by the former are less basis set dependent. 2. Multiple linear regression provides excited state spectra in better agreement with experiment than any simple linear fit that could be performed. 3. Using our best fitting procedure, we obtained a mean absolute error of 6 nm for a set of 66 anthraquinones, with no deviations exceeding 25 nm. The related standard deviation, useful for predictions, is only 8 nm, i.e., $\lambda_{max}^{theo} = \lambda_{max}^{exp} \pm 8$ nm (or ± 0.05 eV) for unknown anthraquinone compounds.

I. Introduction

Today, the molecular modelization techniques offer a competitive alternative for the interpretation of experimental data arising from both academic and industrial measurements. Schäfer recently stated that, in the majority of cases, IR or Raman spectra can be accurately computed with the help of quantum mechanical methods that could be found in many computational chemistry packages.[1] However, it is not the case for UV/VIS spectra of large conjugated molecules. For instance, semiempirical methods, though especially tailored for, are often found to be lucky either inaccurate when reproducing spectral patterns or trends.[1] One of the main difficulties in determining the color of organic compounds is the astonishing accuracy of the standard human eye, which can distinguish, in some parts of the visible spectra (typically in the green region), differences of coloration corresponding to less than 1 nm λ_{max} shifts. Nevertheless, in regard to practical industrial applications, the theoretical calculations could be regarded as serious competitors to experimental approaches for developing new dyes and/or pigments if they were to deliver an estimate of the λ_{max} values within a 5−15 nm accuracy (∼0.05 eV). Such a chemical accuracy for large conjugated molecules is still a tremendous challenge for the modelization approaches. On one hand, highly correlated methods, such as EOM-CC or MR-CI, are completely out of computational reach for molecules possessing several π-electrons and used in solution. On the other hand, as stated above, semiempirical methods are not able to consistently deliver quantitative λ_{max} of dyes. For instance, Adachi and Nakamura reported large errors and poor correlation coefficients with CNDO/s and INDO/s methods for a large set of dyes.[2] In fact, the most promising scheme for systematically evaluating the color of conjugated compounds is the

* Corresponding author e-mail: denis.jacquemin@fundp.ac.be; URL: http://perso.fundp.ac.be/∼jacquemd.

† Research Associate of the Belgian National Fund for Scientific Research.

‡ Facultés Universitaires Notre-Dame de la Paix.

§ Centre technologique FUNDP.

10.1021/ct050287w CCC: $33.50
Published on Web 02/07/2006

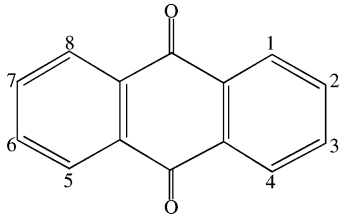


Figure 1. Sketch of anthraquinone with the numbering of substitution positions.

time-dependent density functional theory (TD-DFT).[3] Indeed, TD-DFT is often found robust and efficient for evaluating the low-lying excited spectra of conjugated molecules[4−7] and has been the subject of numerous applications.[8−19] Nevertheless, recent TD-DFT determinations of the λ_{max} of dyes or related conjugated molecules often report mean absolute errors (MAE) in the 0.1−0.4 eV range, i.e., at least twice as large as our target. Indeed, for a large set of sulfur-containing compounds, Fabian obtained MAE of 0.24 eV;[20] the same MAE can be determined for the first $\pi \rightarrow \pi^*$ singlet excitation of the 11 thiouracil for which measurements are given in ref 21. The typically reported TD-DFT error ranges from 0.2 to 0.5 eV for coumarins,[22] from 0.3 to 0.4 eV for uroanic-acid-based molecules,[23] from 0.1 and 0.3 eV for alkyl-amino-benzonitrile compounds,[24] and from 0.3 to 0.4 eV for transition-metal complexes.[25] In 2005, we have found only three published studies with (almost-)quantitative absorption spectra of several dyes. The first is due to Hommen-de-Mello and co-workers who obtained a MAE of 19 nm, for six cationic dyes in water using a PCM-ZINDO// PCM-B3LYP (PCM: Polarizable Continuum Model, see below) approach.[26] The second and third investigations, on diazonium salts and thioindigo, respectively, reported MAE of 6 and 10 nm, using PCM-PBE0 approaches combined with extended basis sets.[18,27] These comparisons highlight the fundamental importance of including solvent effects when simulating absorption spectra of organic dyes.

Two classes of chromophoric unit are principally used as industrial dyes: the N═N chromophore of azo pigments and the C═O group. The latter is present under a variety of chemical forms: coumarins, naphthaquinones, quinacridones, perinones, indigoids ... The carbonyl dyes owe their success to their ability to provide a wide range of colors covering the entire visible spectrum and to their capacity to show long wavelength absorption bands when combined with relatively short π-conjugated systems. In particular, the 9,10-anthraquinones derivatives (Figure 1), in which the central ring bearing two carbonyl groups is fused to two fully aromatic six-member rings, can give rise to a complete range of shades (especially in the green/blue region), depending on the nature and relative position(s) of the auxochromic group(s) substituting hydrogen atom(s) on the outer rings.[28,29] Consequently, anthraquinoidic derivatives represent about 30% of today's world dye production.[28]

Following our first investigation,[30] we aim at setting up an approach able to accurately predict ab initio the λ_{max} of absorption of 9,10-anthraquinones. In addition, we want to assess the basis set effects as well as the relative accuracy of two selected functionals (see below). As experimental input, we have chosen the measurements of Labhart,[31] who obtained the electronic excitation spectra of a large number of anthraquinones, in dichloromethane, with absorption wavelengths almost covering the entire visible spectrum (from 325 to 645 nm). In the Labhart set, one finds a large variety of side groups: hydroxy, amino, nitro, chloro, ... This allows consistent comparisons and meaningful statistical treatment.

This paper is divided as follows. Section 2 gives a description of the quantum-chemical and statistical tools. In section 3.1, we compare the respective accuracy of B3LYP and PBE0 functionals for evaluating the λ_{max} of anthraquinones, whereas in section 3.2, we combine both to obtain an optimal accuracy.

II. Methodology

A. Quantum-Mechanical Calculations. We have chosen the Gaussian03[32] package of programs to perform the geometry optimizations, vibrational analysis, and excited-state evaluations.

The ground-state geometry of each molecule has been fully optimized until the RMS residual force is smaller than 1×10^{-5} au (TIGHT threshold in Gaussian). For anthraquinones, it turns out that the B3LYP[33] functional gives geometries in good agreement with second-order Møller−Plesset structures;[30] B3LYP has therefore been selected. In this functional the exchange is a combination of 20% HF exchange, Slater functional, and Becke's GGA correction,[34] whereas the correlation part combines VWN and LYP[35] functionals. Following each optimization, the vibrational spectrum has been determined at the same level of theory, and it has been systematically checked that all vibrational frequencies are real.

TD-DFT[3] methodology is then used to compute the low-lying excited states of anthraquinone derivatives. We have used two hybrid functionals: B3LYP and PBE0.[36,37] PBE0 is built on the Perdew-Burke-Ernzerhof pure functional,[38] in which the exchange is weighted (75% DFT/25% HF) accordingly to purely theoretical considerations.[39] As expected for this type of dye, the electronic excitation responsible for the color of anthraquinone presents a typical $\pi \rightarrow \pi^*$ character often associated with a large oscillator force. Our theoretical λ_{max} are always related to a transition toward the first singlet excited state, except for some of the dyes absorbing in the 320−350 nm region, for which the experimentally reported λ_{max} corresponds to a higher excited state (but the oscillator strength toward the lower-lying excited states is negligible in that case).

In ref 30, we show that the solvent effects on the ground-state geometry are negligible due to the rigidity of the anthraquinone core but are sizable for transition energies. Therefore, the bulk solvent effects are evaluated during the TD-DFT calculations by means of the standard Polarizable Continuum Model (IEF-PCM).[40,41] In PCM, one divides the problem into a solute part (anthraquinone) lying inside a cavity and a solvent part (in this case, dichloromethane) represented as a structureless material, characterized by its macroscopic properties (dielectric constant, radius, density, molecular volume, ...). PCM is able to obtain a valid approximation of solvent effects as long as no specific interactions (such as hydrogen bonds, ion pairing, ...) link

the solute and the solvent molecules. Because we study UV/Vis spectra, we have selected the so-called nonequilibrium PCM solutions.[41] Indeed the absorption process presents a short characteristic time. Therefore, only the solvent electronic distribution can "adapt" to the new (excited) electronic structure of the solute, while molecular motions of the solvent are frozen during the process.[41]

We use two different atomic basis set combinations in this study. In the less demanding approach (**M-I**), only the Pople's polarized double-ζ basis set, 6-31G(d,p), is selected as our methodological study shows that this basis set could be sufficient.[30] This means that **M-I** corresponds to a PCM-TD-[B3LYP/PBE0]/6-31G(d,p)//B3LYP/6-31G(d,p) approach. In the second method (**M-II**), we use the atomic basis sets often recommended for UV/vis investigations with TD-DFT,[42] i.e., a triple-ζ basis set for the geometry, with additional diffuse functions for the excitation spectra, i.e., **M-II** is PCM-TD-[B3LYP/PBE0]/6-311++G(d,p)//B3LYP/6-311G(d,p).

B. Statistical Treatment. To reach the best agreement between theory and experiment, the results of different approaches can be advantageously combined. To obtain the most efficient combination, the Multiple Linear Regression (MLR),[43−45] which is based on the numerical technique of least-squares fitting and analyzes the relationship between one dependent variable (experimental value) and one or more independent variables (theoretical values), is a method of choice. MLR is a tool for determining the (experimental) property, *Y*, as a function of *p* independent (theoretical) variables (*x*):

$$Y = b_0 + b_1x_1 + b_2x_2 + ... + b_px_p + R \quad (1)$$

To test the significance of a regression curve, the total sum of squares (TSS) is split into two components, the model sum of squares (MSS) and the residual sum of squares (RSS)

$$\mathrm{TSS} = \mathrm{MSS} + \mathrm{RSS} \quad (2)$$

$$\sum_{i=1}^{n}[y_i - \bar{y}]^2 = \sum_{i=1}^{n}[y(x_i) - \bar{y}]^2 + \sum_{i=1}^{n}[y_i - y(x_i)]^2 \quad (3)$$

with y_i the experimental value, $y(x_i)$ the regression value, $\bar{y}$ the average, and *n* the number of points (number of dyes considered). The correlation coefficient reads

$$R^2 = 1 - \frac{\mathrm{RSS}}{\mathrm{MSS}} \quad (4)$$

If the fitted curve passes through all the original data points, the MSS is equal to the TSS and the RSS is zero. In that case $R^2 = 1$. To test the significance of the regression, test calculations are carried out in a so-called analysis of the variance table (ANOVA), where the mean squares MMS and RMS are obtained by reporting MSS to the number of independent variables, and dividing RSS by *n*-*p*-1, respectively. Indeed, R^2 could be abnormally large if there are numerous descriptors (i.e. large *p*) but a few data points (i.e. small *n*). Therefore, one uses an adjusted coefficient

$$R^2_{\mathrm{adj}} = 1 - \frac{n-1}{n-p-1}\left(1 - \frac{\mathrm{RSS}}{\mathrm{MSS}}\right) \quad (5)$$

If the MMS/RMS ratio is significantly large, the null hypothesis may be rejected and the regression is meaningful. Confidence limits for the regression parameters b_i, measuring the adequacy of each independent variables in the model, are also determined. The ratio between b_i and the associated error could be compared to the critical values for which the probability (*P*-value) to obtain a regression coefficient by chance has been tabulated. If necessary, this treatment allows for eliminating step-by-step the less significant independent variables. MLR provides not only the usual mean absolute error (MAE) but also a standard deviation, d_R, computed as

$$d_R = \sqrt{\frac{\mathrm{RSS}}{n-p-1}} \quad (6)$$

d_R is useful for the prediction of the properties of compounds not included in the training set. In the present study, MLR has been performed with the Statgraphics Plus 5.1. program.[46]

III. Results and Discussion

A. Comparison between B3LYP and PBE0. The computed λ_{max} for 66 anthraquinones are reported (in nm) in Table 1 and are compared to the experimental data taken in ref 31. Before using multilinear regression, it is worth evaluating the performance of the two functionals using "raw" values directly extracted form TD-DFT calculations. When using the most accurate theoretical level (**M-II**), we obtain, for B3LYP a MAE of 20 nm (RMS of 25 nm), whereas for PBE0 the MAE is 14 nm (RMS of 18 nm). The corresponding MAE (RMS) in eV are 0.12 (0.16) and 0.08 (0.10) for B3LYP and PBE0, respectively. Thus, both functionals are quite efficient for anthraquinones: the MAE are clearly in the lower range of the expected TD-DFT deviations (see Introduction). This is well illustrated in Figure 2, where the qualitative and quantitative agreements between theory and experiment is striking. In general, B3LYP slightly overshoots the λ_{max} of dyes with large excitation energies, PBE0 presenting the (completely) opposite behavior.

In addition, one can state that PBE0 is statistically more efficient than B3LYP for evaluating the UV/Vis spectra of anthraquinones at a 99% confidence level. This assertion is confirmed by the extreme deviations that are smaller with PBE0 functional: +63/−32 nm for B3LYP and +38/-45 nm for PBE0. Using the less demanding computational scheme, **M-I**, the errors calculated with PBE0 are almost unchanged: MAE = 13 nm (0.08 eV) and RMS = 17 nm (0.09 eV) (in fact there is a 30% probability that the basis set modification does not statistically alter the results obtained with the PBE0 functional). The same is true for the largest deviations: +33/−49 nm. This illustrates that the PBE0 λ_{max} are already (almost) converged with 6-31G(d,p) basis set. Using **M-II** does not reduce the average errors and is therefore quite useless. Of course, this is a general conclusion, and an individual compound might be significantly affected by the basis set change (NMe_2 groups for instance). Nevertheless, the absolute average change when shifting from **M-I** to **M-II** is small: 4 nm (0.03 eV) with PBE0, i.e., significantly smaller than the theory-experiment discrepancies.

For B3LYP, the changes induced by the basis set effects are larger. Indeed, with **M-I** one obtains a MAE of 16 nm

Table 1. Comparison between the Experimental and Theoretical λ_{max} for Anthraquinones in CH_2Cl_2[a]

compound		M-I			M-II			
substitution	no.	B3LYP	PBE0	MLR	B3LYP	PBE0	MLR	expt
2-F	13	327.7	316.4	324.1	334.7	322.1	321.7	325
	47	327.2	315.8	322.9	333.1	320.8	321.4	327
2-Cl	8	331.0	319.4	326.2	336.7	323.8	322.5	330
2,3-Cl	12	335.4	323.6	330.1	339.7	327.1	327.5	330
2,3-Br	15	337.2	324.9	329.3	343.3	329.5	325.6	330
2,6-Cl	72	334.1	321.8	325.7	341.1	327.5	324.1	330
2,7-Cl	73	322.7	311.5	318.9	328.0	316.0	317.0	330
1-NO_2,4-Cl	44	351.6	338.0	338.5	356.5	341.9	336.9	335
1-Cl	7	346.6	334.0	338.5	355.4	341.3	338.1	337
1,8-Cl	11	353.9	340.7	343.5	360.9	346.4	342.5	344
1,5-Cl	9	358.1	344.4	343.5	364.3	349.3	343.9	347
1,4-Cl	10	359.6	346.4	350.1	368.7	353.0	345.5	350
2-OMe	69	398.8	382.8	379.4	406.6	394.6	408.3	363
2-OH	68	395.5	379.7	376.7	400.1	383.6	377.9	365
1-OMe	6	393.9	379.7	384.0	402.1	386.5	384.8	380
1,8-OMe	35	405.6	390.1	390.1	412.5	395.6	390.3	385
1-OH	1	410.9	396.4	402.0	411.2	396.6	400.5	405
1-Cl,2-NH_2	61	449.9	430.8	420.7	458.5	438.2	426.5	405
2-NH_2,3-Br	63	446.3	426.6	413.1	454.6	433.4	417.3	406
1-NO_2,2-NH_2	55	462.3	439.9	416.0	470.9	446.0	417.5	410
1-NHCOMe	65	430.7	415.0	418.0	430.9	414.9	416.2	410
2-NH_2	17	455.1	435.3	422.7	464.8	443.4	428.2	410
2-NH_2,3-Cl	62	444.0	424.7	412.7	453.7	432.9	418.3	414
1-NHCOPh	48	434.1	417.2	415.1	436.2	418.8	415.3	415
1,2-OH	5	439.9	422.8	420.6	437.9	421.0	419.8	416
2-NH_2,3-NO2	56	433.5	417.3	418.4	442.6	425.0	421.7	420
1,5-OH	3	433.2	417.8	422.7	432.9	417.1	419.6	428
1,8-OH	4	435.4	420.2	426.4	437.7	420.1	416.0	430
1-SMe	36	449.8	431.0	422.3	469.0	446.2	426.0	438
2,3-NH_2	19	498.0	475.0	453.9	504.6	480.0	458.2	442
1-NH_2,4-NO_2	57	481.7	462.0	454.1	486.9	465.5	453.9	460
1-NH_2,5-OMe	59	478.3	461.1	464.5	481.2	463.2	464.5	460
1-NH_2	16	475.7	459.0	464.4	480.3	462.6	465.0	465
1-NH_2,2-Me	40	473.0	456.8	464.2	477.7	460.6	465.0	465
1-NH_2,4-Cl	37	475.9	460.0	469.2	479.4	463.0	470.6	466
1-NH_2,6-Cl	38	487.6	470.1	473.6	490.7	472.3	473.5	470
2-NMe_2	74	495.8	474.8	462.9	513.8	489.1	468.4	470
1-NH_2,2-Me,4-Br	46	476.1	460.1	468.9	478.2	461.8	469.2	473
1-NH_2,2-NHCOPh	64	476.7	459.1	460.4	480.6	462.4	462.8	475
1,4-OH	2	472.8	459.2	478.9	467.2	454.1	473.2	476
1-NH_2,6,7-Cl	39	493.1	475.6	480.0	493.9	475.8	478.8	477
1,2-NH_2	21	487.2	469.6	472.5	497.7	475.0	459.8	480
1,5-NH_2	20	485.4	468.6	475.1	490.1	472.2	475.4	480
1,4-NHCOPh	49	507.4	489.4	493.6	507.4	489.2	494.0	490
1,8-NH_2	22	512.6	493.8	495.1	517.7	497.5	495.8	492
1-NH_2,4-OMe	41	505.7	491.0	510.7	508.8	493.6	510.9	500
1-NMe_2	24	516.6	497.8	499.7	531.3	509.5	503.5	504
1-NHMe	23	500.9	483.4	489.0	509.6	490.4	491.4	508
1-NHPh	26	545.3	519.3	491.5	550.2	522.6	495.9	508
1-NHMe,4-Br	66	504.2	487.2	495.7	509.8	491.6	496.8	510
1-NO_2,4,5,8-OH	33	507.9	491.8	505.2	501.6	485.6	498.4	510
1-OH,4-NH_2	25	520.4	506.5	532.2	517.6	503.8	528.2	520
1-OH,2,4-NH_2	43	529.2	513.0	529.3	527.7	511.6	528.2	530
1-NH_2,4-NHCOPh	50	529.6	513.8	532.0	529.6	513.6	531.0	532
1-NHMe,4-OMe	32	530.1	515.1	537.2	537.1	521.2	540.2	540
1,4-NH_2	18	539.8	527.2	562.3	539.0	526.3	558.7	550
1,4-NH_2,2-OMe	60	532.2	519.0	550.0	532.5	519.2	548.1	550
1-OH,4-NHPh	28	581.2	560.4	563.0	580.6	559.1	562.3	566
1-NH_2,4-NHPh	29	585.2	568.3	590.1	586.3	569.0	590.3	590
1,5-NH_2,4,8-OH	45	577.8	563.1	594.2	574.2	559.1	587.4	590
1-NH_2,4-NHMe	31	561.6	548.9	587.0	563.7	550.7	585.9	590
1,4,5,8-NH_2	67	602.8	587.8	621.4	602.4	586.7	617.1	610
1,4-NHMe	34	583.4	570.7	612.2	588.5	575.3	613.7	620
1,4-NHPh	27	621.5	602.0	617.2	623.2	603.1	618.9	620
1-NHMe,4-NHPh	30	606.1	589.2	614.3	609.7	592.3	617.0	625
1,4-NH_2,2-NO_2	58	682.9	657.5	654.3	692.0	662.6	651.6	645

[a] Experimental λ_{max} and anthraquinone reference numbers have been taken from ref 31. The MLR λ_{max} are calculated with eqs 15 and 16. All values are in nm.

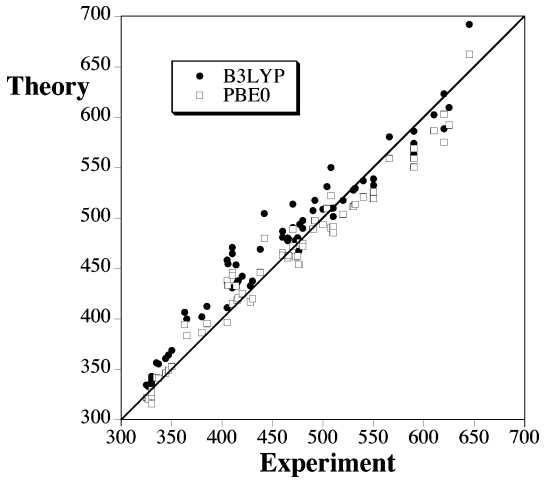


Figure 2. Comparison between the experimental and theoretical (**M-II**) λ_{max}. All values are in nm.

(0.09 eV) and a RMS of 20 nm (0.13 eV), surprisingly significantly (probability > 99%) smaller than with **M-II**. This means that, on one hand, the convergence with basis set size is slower with B3LYP than with PBE0 and, on the other hand, that the use of **M-I** leads to a "lucky" agreement in the case of B3LYP, i.e., there is some functional/basis set error compensation. If this is quite disappointing from the computational chemist point of view, this is useful in practice, as one could select less demanding methods ... to obtain more accurate λ_{max} for the average anthraquinone structure. In section 3.2, we demonstrate that **M-I** is also preferable after statistical treatment.

The substituents leading to significant auxochromic shifts belong to various chemical classes, but, in practice, the major groups used for anthraquinoidic dyes are the hydroxys (and the corresponding OR) present in naturally occurring quinones, and the amines (and NHMe, NMe_2, NHPh, NHCOPh, NHCOMe, ...) generally found in synthesized structures. Both groups are strongly electroactive and often (positions 1, 4, 5, and 8) affect the carbonyl chromophore by internal hydrogen bonds, especially strong with the NHPh groups. For the hydroxy group,[47] selecting **M-I** (**M-II**) we obtained a MAE of 10 (11) nm with PBE0, the corresponding B3LYP figures being 17 nm (22 nm). For the amino auxochroms,[47] the **M-I** (**M-II**) MAE are 18 nm (18 nm) with PBE0 and 18 nm (21) nm with B3LYP. This confirms the better behavior of the PBE0 functional with respect to basis set size. In addition, one clearly sees that the difference between the two functionals is larger when OH and OR groups are grafted to the anthraquinone core. In that case, PBE0 is clearly preferable.

B. Statistical Treatment. If one looks for a predictive tool for determining the color of anthraquinone dyes, the "raw" estimates of TD-DFT can be improved by using statistical treatment, either simple linear regression (SLR) if one uses only one functional, or MLR if the excitation energies obtained with both functionals are combined. The following SLR equations are obtained with nm units

$$\lambda_{max,nm} = -28.80 + 1.040\,\lambda_{max,nm}^{B3LYP-\mathbf{M-I}} \quad (7)$$

$$\lambda_{max,nm} = -24.18 + 1.067\,\lambda_{max,nm}^{PBE0-\mathbf{M-I}} \quad (8)$$

$$\lambda_{max,nm} = -37.64 + 1.049\,\lambda_{max,nm}^{B3LYP-\mathbf{M-II}} \quad (9)$$

$$\lambda_{max,nm} = -33.67 + 1.080\,\lambda_{max,nm}^{PBE0-\mathbf{M-II}} \quad (10)$$

The corresponding relationships in eV read

$$\lambda_{max,eV} = -0.051 + 1.045\,\lambda_{max,eV}^{B3LYP-\mathbf{M-I}} \quad (11)$$

$$\lambda_{max,eV} = -0.036 + 1.003\,\lambda_{max,eV}^{PBE0-\mathbf{M-I}} \quad (12)$$

$$\lambda_{max,eV} = -0.116 + 1.081\,\lambda_{max,eV}^{B3LYP-\mathbf{M-II}} \quad (13)$$

$$\lambda_{max,eV} = -0.098 + 1.034\,\lambda_{max,eV}^{PBE0-\mathbf{M-II}} \quad (14)$$

A statistical analysis of the results obtained with these equations is given in Table 2. PBE0 equations are always more efficient that their B3LYP counterparts, systematically giving larger R^2 and smaller MAE and d_R, as well as a slightly smaller number of extreme deviations. In addition, for equations in the more physical energetic scale, the *a* and *b* [eqs (12) and (14)] are closer to zero and one, respectively [than the corresponding (11) and (13)]. From Table 2, one directly concludes that **M-I** is more appropriate than **M-II** for estimating the λ_{max} of anthraquinone using a simple linear fit. Therefore the SLR results confirm the conclusions of section 3.1: PBE0 has to be selected for TD-DFT calculations on anthraquinone dyes. Using the 6-31G(d,p) basis set, this functional leads to a standard deviation of 15 nm, i.e., the absorption energy of dyes of the same family (but not included in our set) could be predicted with an accuracy of ±15 nm.

Using MLR, the following equations have been obtained:

$$\lambda_{max,nm} = 9.54 - 4.604\,\lambda_{max,nm}^{B3LYP-\mathbf{M-I}} + 5.762\,\lambda_{max,nm}^{PBE0-\mathbf{M-I}} \quad (15)$$

$$\lambda_{max,nm} = -3.29 - 3.922\,\lambda_{max,nm}^{B3LYP-\mathbf{M-II}} + 5.084\,\lambda_{max,nm}^{PBE0-\mathbf{M-II}} \quad (16)$$

$$\lambda_{max,eV} = 0.112 - 5.599\,\lambda_{max,eV}^{B3LYP-\mathbf{M-I}} + 6.350\,\lambda_{max,\,eV}^{PBE0-\mathbf{M-I}} \quad (17)$$

$$\lambda_{max,eV} = 0.036 - 4.260\,\lambda_{max,eV}^{B3LYP-\mathbf{M-II}} + 5.087\,\lambda_{max,eV}^{PBE0-\mathbf{M-II}} \quad (18)$$

In all these equations the *P*-value analysis allows for stating that all DFT coefficients are statistically significant at the 99% confidence level. Figure 3 provides a comparison between the λ_{max} computed with eqs 15 and 16 and experimental results. As can be seen, the agreement is significantly better than in Figure 2, highlighting the interest of such post-treatment of quantum-chemical results. In Table 2, the MLR statistical data are compared to the SLR. With **M-I**, the MLR R^2_{adj} are larger than 99% indicating a nearly perfect fit, the MAE is limited to 6 nm (0.04 eV), the d_R is only 8 nm (0.05 eV), and the maximal deviations are essentially half of those obtained with SLR. It is also striking that none of the 66 estimates exceeds a 25 nm deviation, whereas only 9% present errors larger than 0.1 eV. Using **M-II** leads to slightly poorer R^2_{adj}, MAE, and d_R but to very large errors for 2-OMe, which is clearly a problematic

Table 2. Comparison of the Statistical Parameters Obtained By SLR-B3LYP, SLR-PBE0, and MLR[a]

	M-I		
property	SLR-B3LYP	SLR-PBE0	MLR
R^2 in %	96.0 [96.4]	97.0 [97.2]	99.2 [99.2]
R^2_{adj} in %	96.0 [96.4]	97.0 [97.2]	99.2 [99.1]
MAE in nm [eV]	13.4 [0.079]	11.5 [0.069]	5.8 [0.037]
d_R in nm [eV]	17.7 [0.105]	15.1 [0.092]	7.9 [0.051]
largest positive deviation in nm	47.0 (2,3-NH_2)	40.7 (2,3-NH_2)	16.4 (2-OMe)
largest negative deviation in nm	−42.2 (1,4-NHMe)	−35.2 (1,4-NHMe)	−19.0 (1-NHMe)
largest positive deviation in eV	0.21 (2,7-Cl)	0.20 (2,7-Cl)	0.12 (2,7-Cl)
largest negative deviation in eV	−0.28 (1-NO_2,2-NH_2)	−0.23 (1-NO_2,2-NH_2)	−0.14 (2-OMe)
cases with abs. deviations > 10 nm	34 (52%)	28 (42%)	16 (24%)
cases with abs. deviations > 25 nm	10 (15%)	7 (11%)	0 (0%)
cases with abs. deviations > 0.1 eV	15 (23%)	12 (18%)	6 (9%)
	M-II		
property	SLR-B3LYP	SLR-PBE0	MLR
R^2 in %	94.9 [95.7]	96.3 [96.6]	98.7 [98.2]
R^2_{adj} in %	94.8 [95.6]	96.3 [96.6]	98.7 [98.2]
MAE in nm [eV]	15.2 [0.088]	12.9 [0.077]	6.7 [0.046]
d_R in nm [eV]	19.9 [0.115]	16.9 [0.101]	10.0 [0.074]
largest positive deviation in nm	49.4 (2,3-NH_2)	42.6 (2,3-NH_2)	45.3 (2-OMe)
largest negative deviation in nm	−40.6 (1,4-NHMe)	−32.5 (1,4-NHMe)	−20.3 (1,2-NH_2)
largest positive deviation in eV	0.21 (2,7-Cl)	0.20 (2,7-Cl)	0.14 (2,7-Cl)
largest negative deviation in eV	−0.29 (1-NO_2,2-NH_2)	−0.26 (2-OMe)	−0.39 (2-OMe)
cases with abs. deviations > 10 nm	38 (58%)	38 (58%)	15 (23%)
Cases with Abs. Deviations > 25 nm	13 (20%)	9 (14%)	1 (2%)
Cases with Abs. Deviations > 0.1 eV	20 (30%)	17 (26%)	6 (9%)

[a] These values are obtained with fittings based on the λ_{max} computed in nm and eV.

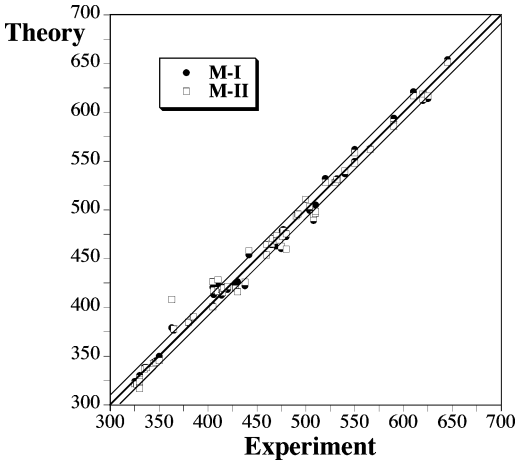


Figure 3. Comparison between the experimental and theoretical λ_{max} obtained by MLR (eqs 15 and 16). The central line indicates a perfect match, whereas the two side lines are borders for ±10 nm discrepancies. All values are in nm.

substitution for our approaches that systematically undershoot the related excitation energy.

IV. Conclusions and Outlook

The absorption spectra of 66 anthraquinone dyes have been computed with a PCM-TD-DFT approach using two hybrid functionals (B3LYP and PBE0) and two basis set combinations. The present study points out that, although both functionals provide at least satisfactory results, PBE0 is more adequate than B3LYP for evaluating the λ_{max} of anthraquinones. In addition, it turns out that the 6-31G(d,p) basis set provides converged transition energies with the PBE0 functional; a further extension of the basis does not improve (and sometimes slightly decreases) the average quality of the theoretical prediction. This means that the absorption spectra of substituted anthraquinones can be accurately evaluated at a relatively small computational cost. We have used a three-step procedure for comparing experimental and theoretical λ_{max}: 1. excitation energies directly taken from TD-DFT calculations, 2. absorption maxima evaluated by SLR, and 3. λ_{max} optimized with MLR using the results of B3LYP and PBE0 functionals. At each step, the accuracy is improved. However, as TD-DFT nicely reproduces the change in λ_{max} resulting from strong auxochromic substitution, the MAE is almost unchanged when using a SLR instead of the "raw" data. For instance, with PBE0, it goes from 13 to 12 nm. The improvement with MLR is more drastic with a MAE limited to 6 nm and a much smaller number of large deviations. More impressively the predicting power of the MLR equations is such that the blind tests for anthraquinones not included in our training set can be estimated with a standard deviation of ±0.05 eV (±8 nm). Although we use a wide panel of substituents, almost covering the entire visible spectrum, the errors reported in this study are much smaller than in most of the recent TD-DFT investigations. This is probably due, in parts, to the explicit consideration of medium effects in our model.

Acknowledgment. E.P. and D.J. thank the Belgian National Fund for Scientific Research for their research associate positions. J.P. acknowledges the FRIA (Belgian "Fonds pour la formation à la Recherche dans l'Industrie et dans l'Agriculture") for his Ph.D. grant. Most calculations

have been performed on the Interuniversity Scientific Computing Facility (ISCF), installed at the Facultés Universitaires Notre-Dame de la Paix (Namur, Belgium), for which the authors gratefully acknowledge the financial support of the FNRS-FRFC and the "Loterie Nationale" for the convention number 2.4578.02 and of the FUNDP.

References

(1) Schäfer, A. *Modern Methods and Algorithms of Quantum Chemistry*, 2nd ed.; volume 3 of *NIC* Johnvon Neumann Institute for Computing: Jülich, 2000.

(2) Adachi, M.; Nakamura, S. *Dyes Pigm.* **1991**, *17*, 287−296.

(3) Runge, E.; Gross, E. K. U. *Phys. Rev. Lett.* **1984**, *52*, 997−1000.

(4) Casida, M. E. In *Accurate Description of Low-Lying Molecular States and Potential Energy Surfaces*; Hoffmann, M. R., Dyall, K. G., Eds.; American Chemical Society: Washington, DC, 2002; Vol. 828.

(5) Onida, G.; Reining, L.; Rubio, A. *Rev. Mod. Phys.* **2002**, *74*, 601−659.

(6) Baerends, E. J.; Ricciardi, G.; Rosa, A.; van Gisbergen, S. J. A. *Coord. Chem. Rev.* **2002**, *230*, 5−27.

(7) Maitra, N. T.; Wasserman, A.; Burke, K. In *Electron Correlations and Materials properties 2*; Gonis, N. K., Ciftan, M., Eds.; Kluwer: Dordrecht, 2003.

(8) Jamorski-Jödicke, C.; Lüthi, H. P. *J. Am. Chem. Soc.* **2002**, *125*, 252−264.

(9) Bartholomew, G. P.; Rumi, M.; Pond, S. J. K.; Perry, J. W.; Tretiak, S.; Bazan, G. C. *J. Am. Chem. Soc.* **2004**, *126*, 11529−11542.

(10) Besley, N. A.; Oakley, M. T.; Cowan, A. J.; Hirst, J. D. *J. Am. Chem. Soc.* **2004**, *126*, 13502−13522.

(11) Ciofini, I.; Lainé, P. P.; Bedioui, F.; Adamo, C. *J. Am. Chem. Soc.* **2004**, *126*, 10763−10777.

(12) Improta, R.; Barone, V. *J. Am. Chem. Soc.* **2004**, *126*, 14320−14321.

(13) Rappoport, D.; Furche, F. *J. Am. Chem. Soc.* **2004**, *126*, 1277−1284.

(14) Stich, T. A.; Buan, N. R.; Brunold, T. C. *J. Am. Chem. Soc.* **2004**, *126*, 9735−9749.

(15) Jacquemin, D.; Preat, J.; Wathelet, V.; André, J. M.; Perpète, E. A. *Chem. Phys. Lett.* **2005**, *405*, 429−433.

(16) Chisholm, M. H.; D'Acchioli, J. S.; Pate, B. D.; Patmore, N. J.; Dala, N. S.; Zipse, D. J. *Inorg. Chem.* **2005**, *44*, 1061−1067.

(17) Masternak, A.; Wenska, G.; Milecki, J.; Skalski, B.; Franzen, S. *J. Phys. Chem. A* **2005**, *109*, 759−766.

(18) Jacquemin, D.; Preat, J.; Perpète, E. A. *Chem. Phys. Lett.* **2005**, *410*, 254−259.

(19) Jorge, F. E.; Autschbach, J.; Ziegler, T. *J. Am. Chem. Soc.* **2005**, *127*, 975−985.

(20) Fabian, J. *Theor. Chem. Acc.* **2001**, *106*, 199−217.

(21) Shukla, M. K.; Leszczynski, J. *J. Phys. Chem. A* **2004**, *108*, 10367−10375.

(22) Cave, R. J.; Castner, E. W., Jr. *J. Phys. Chem. A* **2002**, *106*, 12117−12123.

(23) Danielsson, J.; Ulicny, J.; Laaksonen, A. *J. Am. Chem. Soc.* **2001**, *123*, 9817−9821.

(24) Jamorski-Jödicke, C.; Casida, M. E. *J. Phys. Chem. B* **2004**, *108*, 7132−7141.

(25) Petit, L.; Maldivi, P.; Adamo, C. *J. Chem. Theory Comput.* **2005**, *1*, 953−962.

(26) Hommen de Mello, P.; Mennucci, B.; Tomasi, J.; da Silva, A. B. F. *Theor. Chem. Acc.* **2005**, *113*, 274−280.

(27) Jacquemin, D.; Preat, J.; Wathelet, V.; Perpète, E. A. *J. Mol. Struct. (THEOCHEM)* **2005**, *731*, 67−72.

(28) Green, F. J. *The Sigma-Aldrich Handbook of Stains, Dyes and Indicators*; Aldrich Chemical Company, Inc.: Milwaukee, WI, 1990.

(29) Thomson, R. H. *Naturally Occurring Quinones*, 2nd ed.; Academic Press: London, 1971.

(30) Jacquemin, D.; Preat, J.; Charlot, M.; Wathelet, V.; André, J. M.; Perpète, E. A. *J. Chem. Phys.* **2004**, *121*, 1736−1743.

(31) Labhart, H. *Helv. Chim. Acta* **1957**, *152*, 1410−1421.

(32) Frisch, M. J. et al. *Gaussian 03, Revision B.04*; Gaussian, Inc.: Wallingford, CT, 2004.

(33) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648−5652.

(34) Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098−3100.

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

(36) Adamo, C.; Barone, V. *J. Chem. Phys*. **1999**, *110*, 6158−6170.

(37) Ernzerhof, M.; Scuseria, G. E. *J. Chem. Phys.* **1999**, *110*, 5029−5036.

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

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

(40) Amovilli, C.; Barone, V.; Cammi, R.; Cancès, E.; Cossi, M.; Mennucci, B.; Pomelli, C. S.; Tomasi, J. *Adv. Quantum Chem.* **1998**, *32*, 227−261.

(41) Cossi, M.; Barone, V. J. *Chem. Phys.* **2001**, *115*, 4708−4717.

(42) Wiberg, K. B.; Stratmann, R. E.; Frisch, M. *J. Chem. Phys. Lett.* **1998**, *297*, 60−64.

(43) Dagnelie, P. *Statistique théorique et appliquée. Tome 1. Statistique descriptive et bases de l'inférence statistique*; De Boeck and Larcier: Bruxelles and Paris, 1998.

(44) Dagnelie, P. *Statistique théorique et appliquée. Tome 2. Inférence statistique à une et deux dimensions*; De Boeck and Larcier: Bruxelles and Paris, 1998.

(45) Pollard, J. *A Handbook of Numerical and Statistical Techniques*; Cambridge University Press: Cambridge, U.K., 1979.

(46) *Statgraphics Plus 5.1*; Manugistics Inc.: Herndon, VA, U.S.A., 2000.

(47) For these comparisons, we have considered anthraquinones substituted only by the groups of the given family, i.e., mixed substitution (as in anthraquinone 60) have not been considered.

CT050287W

JCTC Journal of Chemical Theory and Computation

Quantum Chemical Calculations of the Influence of Anchor-Cum-Spacer Groups on Femtosecond Electron Transfer Times in Dye-Sensitized Semiconductor Nanocrystals

P. Persson,*[,†,‡] M. J. Lundqvist,[†] R. Ernstorfer,[§] W. A. Goddard III,[‡] and F. Willig[§]

Department of Quantum Chemistry, Uppsala University, Box 518, SE-751 20 Uppsala, Sweden, Materials and Process Simulation Center, Beckman Institute 13974, California Institute of Technology, Pasadena, California 91125, and Hahn-Meitner-Institut, Glienickerstrasse 100, D-14109 Berlin, Germany

Received June 2, 2005

Abstract: Electronic properties of dye-sensitized semiconductor nanocrystals, consisting of perylene (Pe) chromophores attached to 2 nm TiO_2 nanocrystals via different anchor-cum-spacer groups, have been studied theoretically using density functional theory (DFT) cluster calculations. Approximate effective electronic coupling strengths for the heterogeneous electron-transfer interaction have been extracted from the calculated electronic structures and are used to estimate femtosecond electron-transfer times theoretically. Results are presented for perylenes attached to the TiO_2 via formic acid (Pe−COOH), propionic acid (Pe−CH_2−CH_2−COOH), and acrylic acid (Pe−CH═CH−COOH). The calculated electron transfer times are between 5 and 10 fs with the formic acid and the conjugated acrylic acid bridges and about 35 fs with the saturated propionic acid bridge. The calculated electron injection times are of the same order of magnitude as the corresponding experimental values and qualitatively follow the experimental trend with respect to the influence of the different substitutions on the injection times.

1. Introduction

Light excitation of dye molecules that are chemically bound to a semiconductor electrode can lead to heterogeneous electron transfer from the dye to the semiconductor if the excited state of the dye overlaps the semiconductor conduction band energetically.[1,2] The realization that such photoinduced heterogeneous electron transfer constitutes a highly efficient way to achieve charge separation has paved the way for the development of so-called dye-sensitized solar cells.[3,4] The introduction of nanocrystalline titanium dioxide (TiO_2) electrodes by Grätzel and co-workers meant that high device efficiencies could be achieved by taking advantage of its spongelike morphology giving up to a 1000-fold increase in photoactive surface area compared to traditional flat electrodes.[5] Thousands of organic and organometallic dyes have by now been tested in order to optimize the device efficiency.[6]

The desire to design more efficient devices has also spurred considerable interest in the nature of the ultrafast heterogeneous electron-transfer processes itself, with the hope that better control of the interfacial electronic properties will help to develop better devices.[1] The basic structural and electronic properties of a typical interface are illustrated in Figure 1. To ensure long-term stability of the interfaces, the chromophores are functionalized by special anchor groups, such as carboxylate and phosphonate groups, which are capable of forming strong chemical bonds to the semiconductor electrodes.[6] The electronic contact between the chromophore and the semiconductor can be controlled by insertion of so-called spacer groups between the chromophore and the anchor group.[7] The typical photoinduced charge separation is initiated by light excitation of the dye from its

* Corresponding author e-mail: petter.persson@kvac.uu.se.
[†] Uppsala University.
[‡] California Institute of Technology.
[§] Hahn-Meitner-Institut.

10.1021/ct050141x CCC: $33.50
Published on Web 01/10/2006

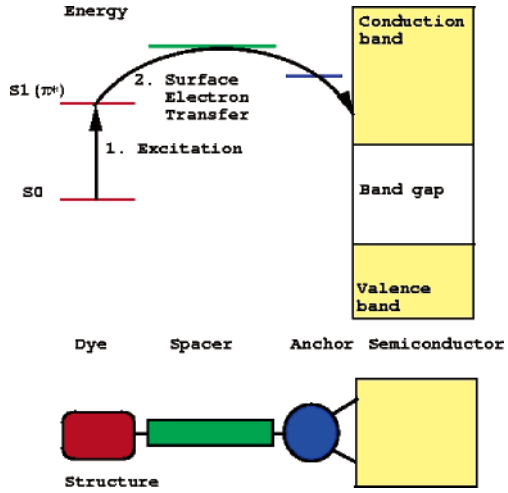


Figure 1. Schematic illustration of the relationship between the molecular structure and the electronic properties of photoinduced heterogeneous electron transfer processes, in dye-sensitized semiconductor devices. The upper and lower panels show the electronic and structural interactions of the various components, respectively. The chromophore part of the dye is attached to the semiconductor via spacer and anchor groups. The photoinduced heterogeneous electron transfer process typically occurs in a two-step process (upper panel), where there is first a local photoexcitation of the dye, followed by surface electron transfer across the spacer-cum-anchor bridge to the semiconductor conduction band which provides a quasi-continuum of electron acceptor states.

electronic ground state to an excited electronic state that is located above the conduction band edge of the semiconductor energetically. The electron injection rate depends strongly on the ability of the anchor-cum-spacer unit that separates the dye from the semiconductor to mediate the electron transfer. In the fastest possible electron transfer processes, the group that anchors the molecule to the semiconductor can act as an efficient conduit of electron transfer by effectively removing the tunneling barrier to the heterogeneous electron transfer.[8]

Ultrafast pump−probe laser spectroscopy has been used extensively to determine electron transfer rates for a number of dye-semiconductor systems with increasingly high time resolution, and it has been shown that electron transfer takes place on a femtosecond time scale in many dye-semiconductor systems.[9−12] The extremely rapid injection rates for a wide range of donors distinguish the heterogeneous electron-transfer processes from most long-range homogeneous, molecular, and biological electron-transfer processes.[13] The enhanced rates for heterogeneous electron transfer are largely caused by the presence of a band of acceptor states offered by the semiconductor substrate, as opposed to the single acceptor state encountered in purely molecular systems, for as long as the donor state lies above the conduction band edge it remains in resonance with a number of acceptor levels.[14] In this situation the electron transfer rate becomes less dependent on vibrational activation compared to most molecular electron transfer reactions that follow the Marcus electron transfer model in which the electron transfer takes place only when the vibrational motion in the donor state reaches a crossing point with the acceptor state where the electronic levels of reactants and products become isoenergetic.[15] The rate of electron transfer is in the heterogeneous case instead largely determined by the strength of the electronic coupling between the excited state of the dye and the semiconductor conduction band. Moreover, the electron transfer in these systems can under favorable conditions take place prior to thermal equilibration of the excited donor state,[16] in contrast to earlier assumptions.[17]

Theoretically, intramolecular and homogeneous electron-transfer processes have been studied extensively.[13,18] Photoinduced heterogeneous electron transfer of dye-sensitized nanoparticles is much less well understood, despite its significant technological potential.[19] Theoretical studies focusing on the conceptual understanding of various aspects of the heterogeneous electron-transfer processes have been presented in the last years.[20,21] The complexity of dye-sensitized semiconductors has, on the other hand, limited the possibilities of quantum chemical calculations and simulations that can provide the predictive power associated with methods that do not rely on fitted parameters.[22] Quantum chemical calculations have been used to investigate the structural and electronic factors involved in the binding of Ru-dye ligands to TiO_2 surfaces as well as studies of the nature of photoinduced electron transfer and charge-transfer excitations of various sensitizers on TiO_2 nanoparticles.[22−35] These studies have, for example, shown that it is important to take both the physical and electronic structure of the interface into account, to model the interface behavior accurately. Structurally, the adsorption can cause significant distortions of the adsorbate structure to accommodate the most favorable surface binding.[22,23] Electronically, the adsorption induces changes in the electronic structure of the adsorbate, in particular in the anchor group.[22,25] Also, the ability to study femtosecond electron-transfer processes from dye molecules to TiO_2 surfaces, either from an electronic coupling,[27,33] a nonadiabatic molecular dynamics,[28] or an electron dynamics[30] perspective, has been explored. The inclusion of nuclear motion in the nonadiabatic molecular dynamics approach makes it particularly attractive for cases where the electronic coupling varies significantly as a result of the motion of the nuclear positions, e.g. due to molecular vibrations.[28]

As reviewed by Noguera, there have been numerous theoretical studies of titanium oxide clusters,[36] including e.g. early work by Bredow and Jug on large TiO_2 anatase particles.[37] To explicitly account for the complexity encountered by using nanocrystalline TiO_2 electrodes, we have also recently made a more systematic computational investigation of small TiO_2 nanocrystals.[38] This approach makes it possible to investigate dye-sensitized TiO_2 nanocrystals using first principles DFT methods.[26,32,33] A potential problem with finite clusters is that they can have poorly developed band structures if they are too small, with e.g. an unphysically large finite level spacing in the substrate bands compared to the experimental situation. Here we use a model TiO_2 nanocrystal with 2 nm diameter which ensures that the splitting of the electronic levels in the relevant part of the conduction band is of the order of 10−20 meV. This spacing

provides an effective lower limit on observable coupling strengths for the dye-semiconductor interaction. A second frequent objection against using clusters as surface models is termination problems requiring the use of surrounding point charges or saturators. In the case of nanocrystals, however, a realistic description of the system may well require the presence of various types of surface motifs catered for by using a sufficiently large finite cluster. From a computational point of view, an advantage of using a cluster approach is that it is relatively unproblematic to accommodate large adsorbates. Periodic calculations would have an intrinsic advantage in automatically providing continuous substrate bands suitable to model the electronic structure of large nanocrystals. In periodic calculations, however, large adsorbates often require the use of very large unit cells in order to avoid undesired interactions across the periodic cell boundaries.

The current work is part of a combined experimental and theoretical investigation of the ultrafast photoinduced electron transfer of a series of perylene derivatives attached to nanostructured TiO_2 through a series of different anchor-cum-spacer groups.[39,40] The lack of spectral overlap of the absorption spectra of the ground-, excited-, and charge-separated states of perylene−TiO_2 interfaces makes it ideal for pump−probe spectroscopic investigations of the fundamental processes involved in photoinduced heterogeneous electron transfer.[16] As a purely organic chromophore attached to TiO_2 via a single anchor group, perylene is also a good candidate for quantum chemical calculations. This system therefore offers exceptionally good opportunities to make direct comparisons between experimental and calculated properties. In this paper we focus on calculated electronic structure properties and the ability to predict heterogeneous electron-transfer rates from first principles density functional calculations.

2. Method

2.1. Electronic Structure Calculations. The properties of the Pe, Pe−COOH, Pe−CH_2−CH_2−COOH, and Pe−CH═CH−COOH molecules, shown in Figure 2, were studied using the B3LYP hybrid functional and the standard 6-31G-(d,p) basis set in the Gaussian03 program.[41] Geometries were fully optimized, and the delocalization of the chromophore HOMO and LUMO orbitals into the anchor-cum-spacer groups was investigated. The S0→S1 excitation was investigated using time-dependent DFT (TD-DFT) calculations with the same functional and basis set. TD-DFT calculations typically give excitation energies accurate to a few tens of an eV for the lowest energy valence excitations at moderate computational cost.[42] There is a danger of spurious low energy excitation energies for charge transfer states due to incorrect treatment of self-interactions when applying TD-DFT to extended systems.[43] TD-DFT calculations have, however, been used successfully to investigate excitations in polycyclic aromatic molecules such as unsubstituted perylene,[44] and the approach appears to work well in the cases where the lowest valence excitations do not have strong charge-transfer character.

For the calculations of sensitized nanocrystals, a $(TiO_2)_{60}$ cluster fulfilling the recently suggested requirements of a

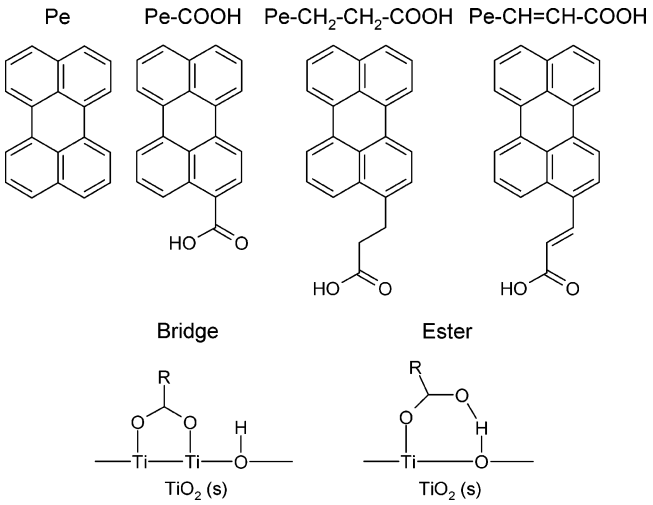


Figure 2. Investigated perylene (Pe) derivatives and surface binding modes. The investigated molecules are perylene (Pe), perylene with formic acid (Pe−COOH), propionic acid (Pe−CH_2−CH_2−COOH), and acrylic acid (Pe−CH═CH−COOH) anchor-cum-spacer groups. The carboxylic acid anchor group was considered to bind to the TiO_2 nanocrystal in a 2M-bidentate (bridge) fashion. For Pe−COOH, binding to the TiO_2 in molecular 1M-monodentate (ester) fashion was also considered.

nanocrystal was used.[38] Briefly, it has been found that neutral, stoichiometric clusters down to approximately 1 nm in diameter constructed so as to accommodate high coordination of all atoms (compatible with their formal oxidation states), and having a small or vanishing dipole moment, display significant structural stability and a well developed band structure.[38] Geometry optimizations of the here investigated systems were performed using DFT calculations with the PW86 exchange functional and the PW91 correlation functional together with a Slater Type Orbital (STO) Valence Single-Zeta (VSZ) basis set and large frozen cores as implemented in the ADF program.[45] This method combination is in the following referred to as PW/VSZ. The local structure of the interface between the TiO_2 nanoparticle and the various sensitizers were optimized at the same level of theory using a smaller $(TiO_2)_5(H_2O)_5$ cluster model. The different sensitizers were optimized on this cluster with relaxation of the local surface environment including the substrate atoms in the vicinity of the adsorbate, while keeping the saturating H_2O as well as the fringe atoms of the substrate cluster fixed. This stepwise optimization approach allows the reconstruction of supramolecular models with optimized adsorbates on optimized nanocrystals, including local relaxation of the nanocrystal in the vicinity of the adsorbate, at an affordable computational cost. It can be noted that water molecules are only used in order to saturate the small cluster used to optimize the adsorbate position on the substrate, and not the large nanocrystal. They are necessary in the small cluster model to saturate unphysical dangling bonds not present in the large nanocrystal which, as described above, has a sufficiently high coordination of every atom to ensure a reasonable description of the effective electronic band structure.[38] To further investigate the influence of the anchor group, a model system was constructed with an unsubstituted

perylene molecule (Pe) placed in the same position relative to the TiO_2 nanocrystal as in the optimized planar bridge binding case. As the aim was to compare the calculated electronic properties with and without anchor groups, rather than the interaction of a physisorbed perylene with a TiO_2 nanocrystal per se, this structure was based on the separately optimized parts without further optimization.

The electronic structure of the combined system was subsequently calculated with B3LYP using a split-valence basis set with large Effective Core Potentials (ECPs) using Gaussian03.[41] In these calculations, all atoms have a Gaussian Type Orbital (GTO) Valence Double-Zeta (VDZ) basis set, except oxygen which has a Valence Triple-Zeta (VTZ) basis set in order to allow a realistic representation of the negative ions in the nanocrystal. This particular method combination is referred to as B3LYP/VD(T)Z in the following and has been used in several previous investigations with good results to describe the electronic structure of systems comprising organic adsorbates on TiO_2 surfaces.[22,27] The B3LYP/VD(T)Z electronic structure calculations on PW/VSZ optimized geometries are referred to as B3LYP/VD(T)Z// PW/VSZ. As the electronic properties are more sensitive to the size of the basis set compared to the structural ones, the combination of a relatively small basis set for the optimizations together with a larger basis set for single point calculations of the electronic structure has been shown to offer a viable computational approach for these complex systems.[27] The basis set used here has, in particular, been used with good results for both structural and electronic properties in previous investigations of organic adsorbates on TiO_2 substrates.[22,27] Of particular relevance to the present application is that, as discussed previously,[22,27] this level of theory gives a reasonable calculated band structure of the TiO_2 substrate. A more detailed investigation of the structural and electronic properties of pure TiO_2 nanocrystals is underway and will be presented in due course.

2.2. Analysis of Interfacial Electronic Interaction. The solution to the time dependent Schrödinger equation has previously been calculated for the system under study assuming a constant value for the electronic coupling strength that reproduces the experimental time scale.[20] The time dependent pump−probe signal shows a monoexponential decay independent of the assumed strength for the electronic coupling as long as the molecular donor state is positioned high enough above the bottom of the empty conduction band of the semiconductor (wide band limit). In the latter case the decay behavior is virtually identical to that predicted by the Fermi's Golden rule perturbation treatment even though there is no restriction on the strength of the electronic coupling. This is not valid any more when the molecular donor level shifts closer to the conduction band edge. The validity of the wide band limit for the perylene chromophore and anatase or rutile TiO_2 has been confirmed experimentally with UPS and 2PPE measurements.[39,40]

Here, we consider the photoinduced surface electron transfer process based on evidence from explicit electronic structure calculations. The initial photoexcitation primarily involves exciations from the Highest Occupied Molecular Orbital (HOMO) to the Lowest Unoccupied Molecular Orbital (LUMO) on the perylene chromophore. The HOMO level, loosely corresponding to the perylene ground state, lies energetically in the band gap region of the TiO_2, as illustrated schematically in Figure 1. Although its energy may be shifted when it is adsorbed on the surface, the electronic interaction with the substrate is believed to be weak. This makes it readily identifiable as a single molecular level in an energy diagram, with a negligible broadening. The LUMO levels of all the chemically anchored sensitizers, on the other hand, are expected to show significant interaction with the conduction band, manifested in the splitting of the isolated sensitizer LUMO to a number of mixed sensitizer-semiconductor levels upon adsorption. According to the Newns-Anderson model for adsorbates on surfaces,[46] the effect of the adsorption on a molecular electronic level, i, is characterized by an energy shift, ΔE_i, relative to its gas-phase value, $E_i(\mathrm{g})$, and a lifetime broadening, $\hbar\Gamma_i$. The shift in energy is related to the gas-phase value by

$$E_i(\mathrm{ads}) = E_i(\mathrm{g}) + \Delta E_i \tag{1}$$

The lifetime broadening is described by a Lorentzian distribution that results from the decay of the excited molecular state resonantly coupled to a continuum of final, charge-separated, states.[47]

A detailed analysis of the electronic structure is necessary in order to quantify both the energy shift and broadening. In an attempt to quantify this interaction, a numerical fitting procedure of the Projected Density of States (PDOS) contributions has been implemented. First, an energy interval was selected within which the adsorbate PDOS contributions were considered to belong to the sensitizer LUMO. The interval was selected so that the PDOS contributions within this interval summed to one orbital. Generally this condition could be achieved to within 2%. This approach can only be expected to work in cases, such as this, where the considered orbital contributions are well separated in energy from those of other molecular orbitals. In more complicated cases, it will be necessary to use a more sophisticated approach involving orbital projection schemes to separate the different contributions. For the selected energy range, the calculated orbital energies, ϵ_i, of the combined system were weighted by the PDOS contributions, p_i, to obtain a weighted average calculated energy, $E_{\mathrm{LUMO}}(\mathrm{ads})$. Specifically, a molecular orbital, ψ_i, is expressed as a linear combination of n atomic orbitals, χ_j^A, centered on atom A, and with expansion coefficients c_{ij}^A:

$$\psi_i = \sum_j^n c_{ij}^A \chi_j^A \tag{2}$$

The portion of the orbital located on the adsorbate is taken to be p_i, which is given by the sum of the squares of the atomic orbital coefficients that are located on the adsorbate (ads) atoms.

$$p_i = \sum_j^{A\in\mathrm{ads}} (c_{ij}^A)^2 / \sum_j^n (c_{ij}^A)^2 \tag{3}$$

The inclusion of the denominator in eq 3 ensures that the probability is properly normalized. This is used to overcome potential complications associated with the fact that the atomic orbitals do not form an orthonormal set and that the sum of the coefficients squared is not strictly one. More stringent assignments would be possible, e.g. by taking the overlap matrix into account, but we have not found such a procedure necessary here.

We take the position of the adsorbate LUMO level in the combined system to be given by the weighted average:

$$E_{\mathrm{LUMO}}(\mathrm{ads}) = \sum_i p_i \epsilon_i \tag{4}$$

Subsequently, quantitative measures of the width of the energy distribution of the LUMO contributions were sought from calculated mean deviation (MD) and root-mean-squared (RMS) values of the selected set of PDOS contributions

$$\hbar\Gamma_{\mathrm{MD}} = p_i|\epsilon_i - E_{\mathrm{LUMO}}(\mathrm{ads})| \tag{5}$$

$$\hbar\Gamma_{\mathrm{RMS}} = \sqrt{\sum_i p_i(\epsilon_i - E_{\mathrm{LUMO}}(\mathrm{ads}))^2} \tag{6}$$

To test the accuracy of the assignment of an effective width to a finite distribution of levels, sets of discrete peaks following a Lorentzian distribution were created and analyzed according to the scheme outlined above for a variety of initial line widths. The calculated MD and RMS line widths were both found to yield correct orders of magnitudes and trends for a wide range of discrete Lorentzian distributions, although the quality of the fit depended on the chosen line width and spacing. For a spacing of 20 meV (similar to that found in the $(TiO_2)_{60}$ cluster) and line widths in the 1−150 meV range, the MD analysis yielded calculated line widths that matched the true value to within 20%. Generally the MD analysis was found to underestimate stronger couplings. The RMS analysis was found to be somewhat less robust compared to the MD analysis, with a tendency to overestimate the line widths for small couplings but to be of similar quality or better than the MD results for line widths exceeding 150 meV. In the analysis below, we have used the MD line widths. It can be noted that these simple ways to approximate a broadening of a molecular level in the presence of a substrate band may in future applications be replaced by a direct fit of the adsorbate level distribution to a Lorentzian function.

The results of the line width-fittings are used to construct Lorentzian distributions, ρ_{LUMO}, with width $\hbar\Gamma$ centered at $E_{\mathrm{LUMO}}(\mathrm{ads})$.[46,47]

$$\rho_{\mathrm{LUMO}}(E) = \frac{1}{(E - E_{\mathrm{LUMO}}(\mathrm{ads}))^2 + \left(\frac{\hbar\Gamma}{2}\right)^2} \tag{7}$$

Finally, the ability of the calculated energy broadenings to capture essential features of the electronic coupling, in the wide band limit of heterogeneous electron-transfer encountered here, was considered by using the calculated PDOS broadenings as an effective measure of an electronic coupling strength that can be converted to an electron-transfer time according to[46,47]

$$\tau = \hbar/\hbar\,\Gamma \tag{8}$$

In convenient numerical units this becomes

$$\tau(\mathrm{fs}) = 658/\Gamma(\mathrm{meV}) \tag{9}$$

3. Results

3.1. Molecular Properties. Molecular properties of the substituted perylenes were first investigated using the standard B3LYP/6-31G(d,p) methodology. In particular, the S0→S1 excitation that is responsible for the photoinduced charge-separation was investigated. This excitation is dominated by the promotion of an electron from the perylene HOMO to the perylene LUMO orbital.[44] The HOMO and LUMO molecular orbitals for the four different molecules are shown in Figure 3. The results for the unsubstituted perylene molecule are similar to those published by Halasinski et al.[44] For all substituents, both the HOMO and LUMO orbitals are delocalized π orbitals. The HOMO orbitals of all the different systems are, moreover, very similar to the HOMO orbital of the unsubstituted perylene molecule, with essentially negligible contributions on the anchor-cum-spacer groups. The LUMO orbital of Pe−CH_2−CH_2−COOH is also very similar to that of the unsubstituted perylene, consistent with the notion that the saturated spacer group is a poor mediator of electron delocalization. The Pe−COOH and Pe−CH═CH−COOH molecules, on the other hand, show considerable delocalization of the perylene LUMO orbital into the anchor-cum-spacer moiety. Interestingly, this is in both cases accompanied by a reorganization of the perylene π* part of the orbital, compared to its symmetrical appearance in perylene itself, in such a way that the perylene π* orbital is located to a larger extent in the vicinity of the substituent. In Pe−COOH, the LUMO orbital can be recognized as a bonding combination between the first perylene π* orbital with the first carboxylate π* orbital. In the Pe−CH═CH−COOH molecule, the LUMO orbital is seen to be essentially a bonding combination of the first π* orbitals of each of the individual parts of the chromophore-spacer-anchor system. It is noteworthy that the delocalization across the unsaturated spacer is sufficiently effective for the anchor group π* to carry nearly equal weight to the LUMO as in the Pe−COOH system. Such delocalizations are consistent with facilitated electron transfer across the anchor-cum-spacer unit.[8,25]

3.2. Molecular Excited States. The vertical excitation energies for the S0 → S1 transition has been calculated using TD- B3LYP/6-31G(d,p) for the various anchor-cum-spacer groups. The results are listed in Table 1. The excitation is in all cases dominated by the HOMO−LUMO transition with no significant charge-transfer character. This means that the LUMO delocalization discussed above is reflected also in the electron distribution of the S1 state, giving further support to the notion that the delocalization of the LUMO facilitates interfacial electron injection. The Pe−COOH and Pe−CH═CH−COOH cases that showed the largest delocalization of the LUMO also show a significant red-shift of the absorption of 0.2 and 0.3 eV, respectively. Twisting the plane of the

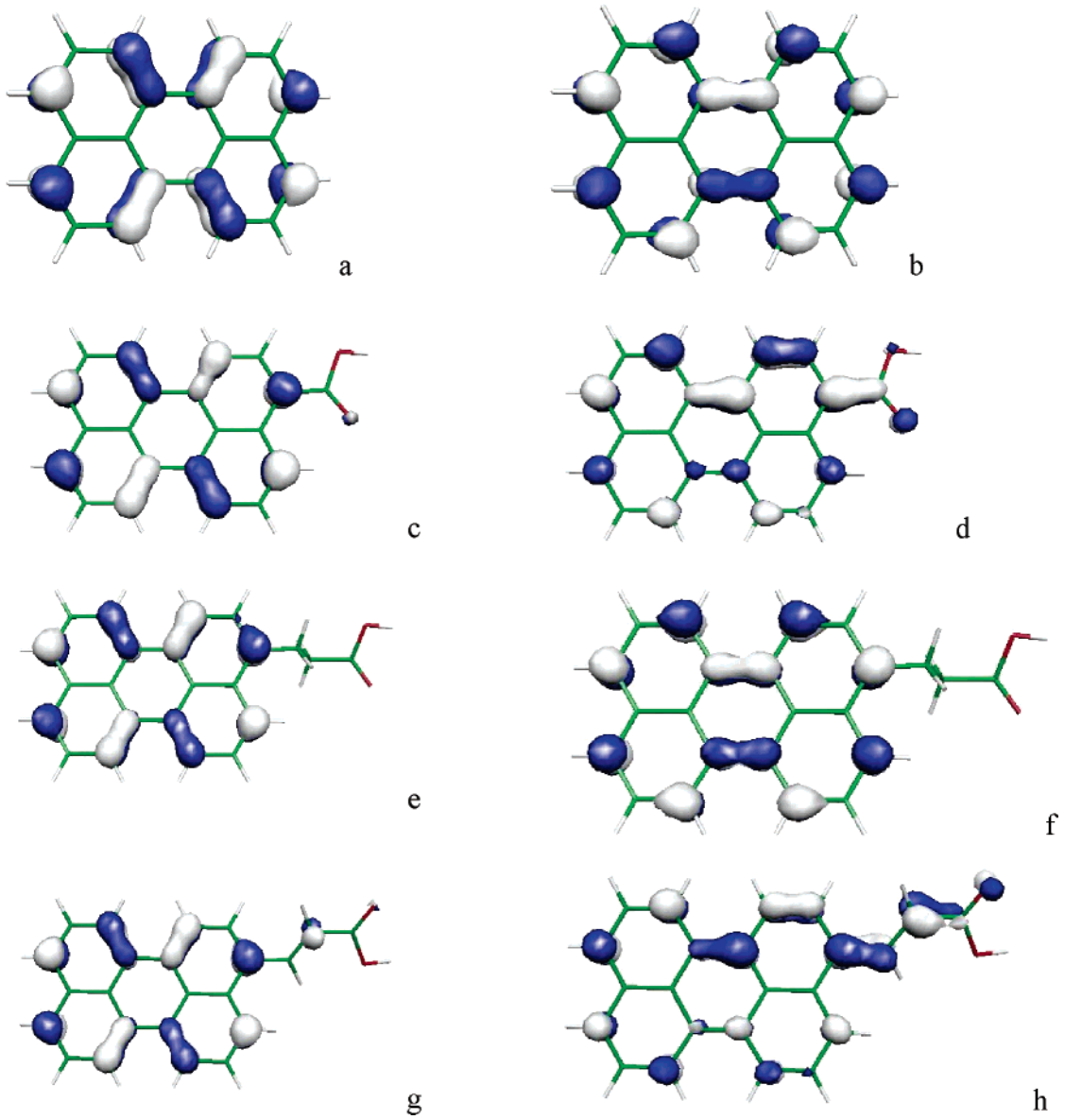


Figure 3. HOMO (left column) and LUMO (right column) orbitals of Pe (a,b), Pe−COOH (c,d), Pe−CH_2−CH_2−COOH (e,f), and Pe−CH=CH−COOH (g,h) according to B3LYP/6-31G(d,p) calculations.

Table 1. Vertical S0→S1 Exciation Energies of Perylenes with Different Anchor-Cum-Spacer Groups, Calculated Using Time Dependent B3LYP/6-31G(d,p)[a]

molecule	*E*/eV	λ/nm	*f*	excitation
Pe	2.89	428.4	0.36	0.62(HOMO→LUMO) −0.11(HOMO-4→ LUMO+2)
Pe−COOH	2.69	461.0	0.40	0.62(HOMO→LUMO)
Pe−CH_2−CH_2−COOH	2.85	435.2	0.43	0.62(HOMO→LUMO)
Pe−CH=CH−COOH	2.60	477.5	0.57	0.62(HOMO→LUMO)

[a] The excitation energies, *E*, wavelengths, λ, oscillator strengths, *f*, and main contributions (excitation coefficients > 0.1) are included in the table.

carboxylate anchor group 90 degrees relative to the perylene plane in Pe−COOH results in a decoupling of carboxylate and perylene π orbitals and a reduction of the red-shift in the absorption compared to the pure perylene case. The injection rates can thus be expected to be sensitive to the detailed structure of the adsorbed chromophore, as already suggested for the isonicotinic acid model chromophore.[27] Furthermore, if the shifts are sufficiently large to be observable spectroscopically, they can serve as a sensitive probe for the local geometry once any adsorption-induced shifts have been taken into account.

3.3. Geometry of the Sensitized Nanocrystals. Atomistic models for the various sensitizers bound to a TiO_2 nanocrystal were constructed from a common, fully optimized, $(TiO_2)_{60}$ cluster. The sensitizer geometries as well as local substrate relaxations in the vicinity of the adsorption site were obtained from geometry optimizations of the sensitizers on a smaller $(TiO_2)_5(H_2O)_5$ cluster. The initial geometry of the smaller cluster was taken from a prototypical anatase (101) surface region of the full $(TiO_2)_{60}$ cluster. To combine a consistent treatment of the sensitized $(TiO_2)_5(H_2O)_5$ and $(TiO_2)_{60}$ clusters, with local surface relaxation near the adsorption site, the atoms at the perimeter of the $(TiO_2)_5$-$(H_2O)_5$ cluster were saturated by hydrogen atoms or hydroxyl groups. The edge atoms together with the hydrogen and

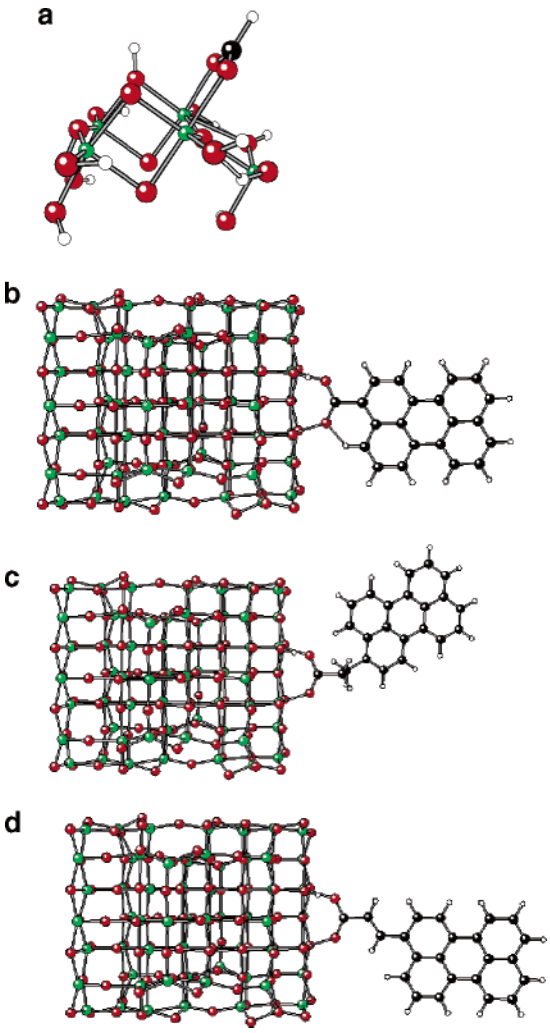


Figure 4. Optimized geometries of sensitized titanium dioxide clusters: (a) HCOOH−$(TiO_2)_5(H_2O)_5$, (b) bridge-binding, planar Pe−COOH−$(TiO_2)_{60}$, (c) Pe−CH_2−CH_2−COOH−$(TiO_2)_{60}$, and (d) Pe−CH═CH−COOH−$(TiO_2)_{60}$. Note that the proton from the carboxylic acid has been transferred to a surface oxygen in the bridge binding mode displayed in 1b−d, in accordance with the bridge binding scheme in Figure 2.

hydroxyl saturators were kept fixed during the subsequent optimization, while the sensitizer and the $(TiO_2)_5(H_2O)_5$ cluster atoms in the vicinity of the adsorption site were fully optimized. Finally, the locally optimized small cluster models were reintroduced into the framework of the large cluster. The geometries of carboxylic acid anchored to the $(TiO_2)_5$-$(H_2O)_5$ cluster and the three combined sensitizer-$(TiO_2)_{60}$ clusters are shown in Figure 4. The PW/VSZ optimized structures agree well with the results of published information about the binding of carboxylic acids to anatase TiO_2 surfaces.[48,49] However, for a nanocrystal such as the $(TiO_2)_{60}$ cluster used here, there are a large number of local adsorption sites with different absorption possibilities. As the present paper focuses on the electronic aspects of the interfaces, we have not tried to find the overall most favorable adsorption site on the nanocrystal. The selected adsorption site should instead only be viewed as a typical surface site, for which we have investigated the effect of the binding on the electronic properties by optimizing both an ester and a bridge binding carboxylic acid, see Figure 2. These two modes are both favorable on TiO_2, and it is not unreasonable to assume that they will either exist in parallel or that the detailed experimental conditions will determine which binding mode prevails. Although beyond the scope of the present paper, more systematic investigations of the binding of anchor groups to nanocrystals, for example comparing different surface sites, investigating the dependence of the size and shape of the nanocrystal, and making comparisons to the binding on surfaces, are interesting and underway.

3.4. Electronic Structure of the Sensitized Nanocrystals. The electronic structures of the combined sensitizer-nanocrystal systems were calculated at the B3LYP/VD(T)Z//PW/VSZ level. An effective total Density of States (DOS) was in each case constructed from the calculated orbital energies using an arbitrary Gaussian broadening of 0.3 eV, and the results are shown in Figure 5. The sensitizer contributions to this DOS have also been extracted using the appropriate atomic orbital coefficients. These contributions are shown in Figure 5 as the 0.3 eV broadened PDOS. Due to the large number of substrate atoms, the total DOS is in all cases dominated by the $(TiO_2)_{60}$ cluster contributions. The calculated DOS is thus very similar for all the studied systems and is only shown as a whole for the nonbound Pe-$(TiO_2)_{60}$. The total DOS spectra display a completely occupied valence band below ca. −7 eV and a completely empty conduction band above ca. −4 eV. The valence and conduction band energies are in all cases within 0.5 eV of the −7.25 and −3.54 eV values calculated for the valence and conduction band edges of an unsensitized $(TiO_2)_{60}$ nanocrystal, respectively. This indicates that the $(TiO_2)_{60}$ cluster model gives a robust and realistic representation of the TiO_2 band gap, both compared to experiment[3] and to periodic TiO_2 calculations using the B3LYP functional with a similar basis set.[27]

The sensitizer contributions to the electronic structure, in the region of interest for the photoexcitation processes involving the chromophore ground and first excited states, can be seen by focusing on the adsorbate PDOS, which are shown in Figure 5 with a magnification of the *y*-axis by a factor of 10 compared to the total DOS for all the different systems. The perylene HOMO π orbital is in all cases easily recognized as a single level located in the band gap region at ca. −6 eV. The LUMO π^* orbital is distributed into a number of contributions to mixed molecule-semiconductor levels around −3 eV. This is about 0.5 eV above the conduction band edge, indicating that heterogeneous electron transfer is energetically possible from the sensitizer LUMO orbital involved in the first excited sensitizer state which was shown earlier to be dominated by a HOMO−LUMO excitation.

3.5. Electronic Coupling Strength. The electronic coupling strength governing the photoinduced heterogeneous electron transfer is likely to be determined largely by the interactions between the sensitizer LUMO orbital and the substrate conduction band. To investigate this interaction more thoroughly, we have made a detailed investigation of the adsorbate LUMO PDOS. The results for the various sensitizers are shown in Figure 6. The figure clearly shows that the calculations predict that the investigated systems

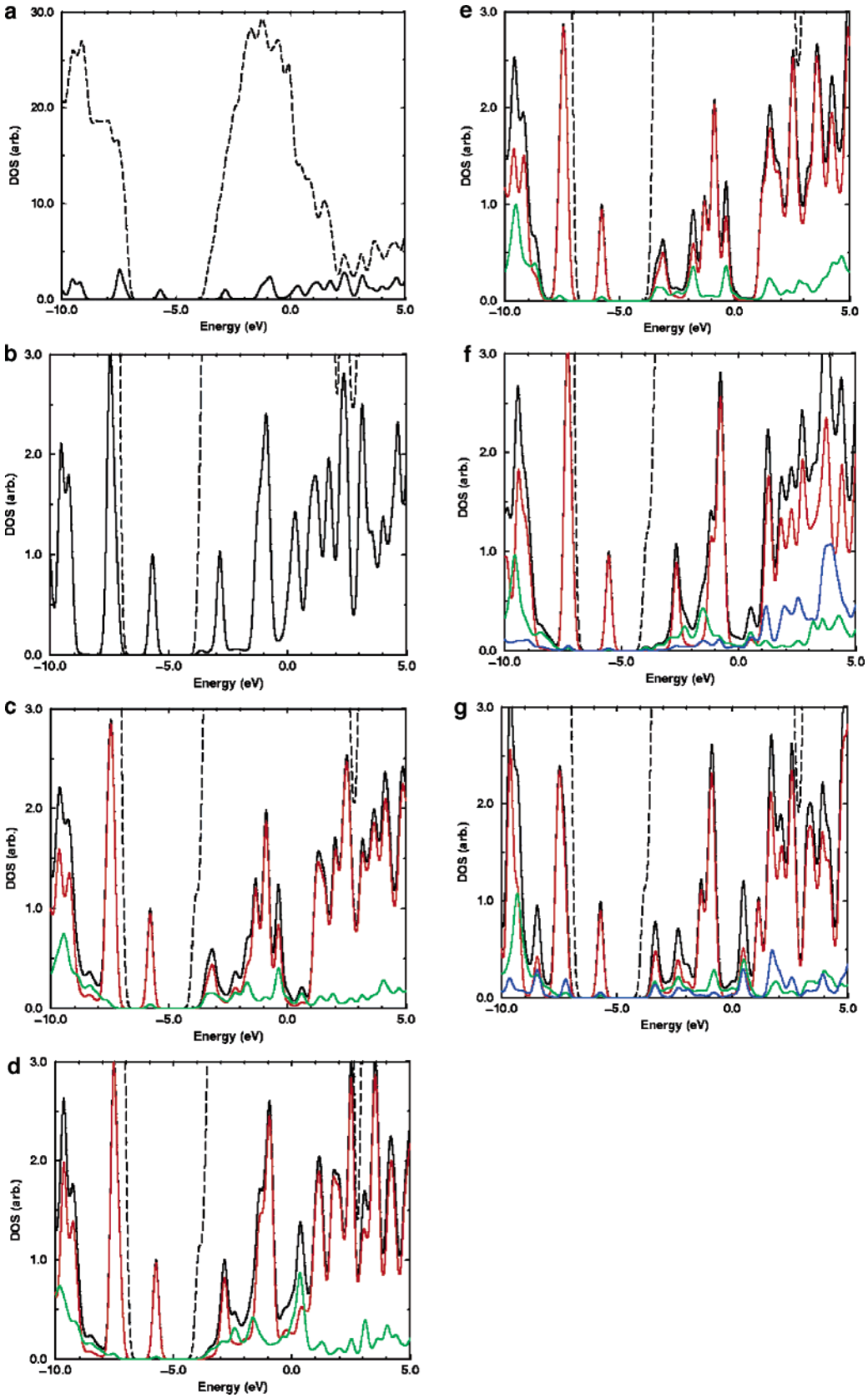


Figure 5. Total and projected DOS plots of the sensitizer-nanocrystal systems. The different panels show (a) Pe−$(TiO_2)_{60}$ DOS, (b) Pe−$(TiO_2)_{60}$ adsorbate PDOS, (c) Pe−COOH−$(TiO_2)_{60}$ adsorbate PDOS for planar bridge adsorption, (d) Pe−COOH−$(TiO_2)_{60}$ adsorbate PDOS for twisted bridge adsorption, (e) Pe−COOH−$(TiO_2)_{60}$ adsorbate PDOS for planar ester adsorption, (f) Pe−CH_2−CH_2−COOH−$(TiO_2)_{60}$ adsorbate PDOS, and (g) Pe−CH═CH−COOH−$(TiO_2)_{60}$ adsorbate PDOS. In all cases: dashed line − total DOS, black line − total adsorbate PDOS, red line − Pe PDOS, green line − COO PDOS, blue line − spacer group PDOS. The DOS and PDOS plots rely on an arbitrary 0.3 eV Gaussian broadening of the calculated orbital energies, used to facilitate visual comparisons.

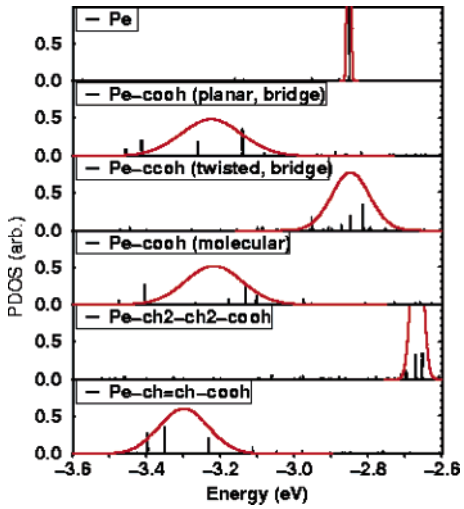


Figure 6. LUMO PDOS plots for the substituted perylenes on $(TiO_2)_{60}$. Black lines − sensitizer orbital (PDOS) contributions. Red curves − Lorentzian fitted curves, ρ_{LUMO}, with parameters from Table 2. To facilitate visual comparison between the PDOS and the Lorentzian curves, the heights of the curves have been scaled so that a curve with a 100 meV fwhm has a height of 0.5.

differ both in the exact position of the LUMO level and the degree to which the sensitizer LUMO orbital mixes with the substrate conduction band. A more detailed analysis is possible from a consideration of the results of the MD-fittings described in section 2.2 of the sensitizer LUMO broadening which are presented in Table 2 and represented graphically in Figure 6 as Lorentzian distributions, ρ_{LUMO}, with width $\hbar\Gamma$ centered at E_{LUMO}(ads). Although the PDOSs obtained from the electronic structures consist of a finite number of states which do not generally follow a simple Lorentzian distribution, visual comparison of the fitted functions with the plot of the individual PDOS contributions in Figure 6 indicates that the fitting successfully captures trends in terms of both energy shifts and broadenings.

The electron-transfer times estimated from the analysis of the quantum chemical calculations fall in the femtosecond time range and are compared with experimental values in Table 2. The listed experimental values were obtained from a monoexponential fit to the measured rise of the molecular product state, i.e., the ionized perylene chromophore, which was monitored as characteristic absorption signal in a femtosecond laser pump−probe experiment. The experiments were carried out with cross-correlation signals of below 25 fs width (fwhm). The perylene dyes were adsorbed from solution on the inner surface of nano-structured anatase TiO_2 layers of about 2 micrometer thickness. The measurements were carried out in ultrahigh-vacuum. Further details are described in a recent Ph.D. thesis by R. Ernstorfer.[39] A very similar trend, with very similar absolute values for the injection times of the same perylene dyes, was measured even more recently applying the technique of femtosecond two-photon photoemission. In the latter case the same perylene dyes were adsorbed but on the (110) surface of a rutile TiO_2 single crystal. Details of these measurements can be found in a recent Ph.D. thesis by L. Gundlach.[40]

As seen in Table 2, the theoretically estimated injection times for the three systems where a direct comparison can be made with experimental measurements are of the right order of magnitude, with τ_{calc} up to a factor of 2 faster than τ_{exp} in all three cases. As the discrepancy between theory and experiment is systematic between the investigated systems, the agreement in terms of the relative injection times is better than in terms of absolute rates. In particular, the calculations predict that the introduction of the saturated ($-CH_2-CH_2-$) spacer slows down the injection by about a factor of 5 compared to the Pe−COOH case, while the corresponding introduction of the unsaturated ($-CH=CH-$) spacer leaves the injection time essentially unaltered. This is in good agreement with the experimental ratios of 4.3 and 0.8, respectively.

The results for the nonbound Pe−$(TiO_2)_{60}$ structure show that without the presence of the anchor group the electronic coupling is reduced substantially. The LUMO of the pure perylene is concentrated almost entirely to a single molecular level at −2.85 eV. This constitutes a significantly weaker coupling compared to all the anchored chromophores. Comparing with the twisted bridge situation, the presence of the anchor group therefore seems to play an important role in enhancing the interfacial electronic coupling also when it is not directly involved in delocalization of the donor level.

There are a number of factors that can influence the calculated absolute injection times. This includes both purely computational effects and discrepancies between the calculated and experimental systems. In terms of the calculations, more work is needed to test the performance of different density functional methods, and basis set effects. Another potential source of error is that it is assumed in the

Table 2. Electronic Interactions between the First Unoccupied Sensitizer Level (LUMO) and the TiO_2 Conduction Band for Systems with Different Anchor-Cum-Spacer Groups and Adsorption Modes[a]

sensitizer	adsorption mode	E_{LUMO}(ads)/ eV	$\hbar\Gamma_{calc}$/ meV	τ_{calc}/ fs	τ_{exp}/ fs
Pe	nonbound	−2.85	2	330	NA
Pe−COOH	planar bridge	−3.22	139	5	13
Pe−COOH	twisted bridge	−2.85	68	10	
Pe−COOH	ester	−3.22	140	5	
Pe−CH_2−CH_2−COOH	bridge	−2.67	20	33	57
Pe−CH=CH−COOH	bridge	−3.30	102	6	10

[a] The table includes the position of the sensitizer LUMO level in the combined system as the weighted average energy, E_{LUMO}(ads), the calculated effective broadening of the level, $\hbar\Gamma_{calc}$, due to the interaction with the surface from the MD analysis described in the text as well as calculated and experimental heterogeneous electron-transfer times.

calculations that electron transfer occurs from the LUMO, whereas in the actual experiment electron transfer occurs from the excited electronic singlet state of perylene. Correspondingly, the energy shifts calculated for the LUMO when the perylene chromophore is attached to the different anchor-cum-spacer groups cannot be seen with the same magnitude for the excited state as is borne out by the absorption spectra and also by the UPS and 2PPE measurements probing the energy of the excited singlet state of the perylene chromophore with respect to the lower edge of the conduction band of the semiconductor TiO_2.[39,40] Dynamic effects could also contribute to the discrepancy, and the optimized geometries used in the present calculations may, in fact, correspond to geometries where the LUMOs are more strongly coupled to the substrate conduction band compared to the average value during the thermal and vibrational motion of the system. In terms of the compatibility with the experimental system, the present calculations assume an ideal surface termination, with direct chemical bonding from the anchor group to substrate Ti atoms. If the experimental systems are not atomically clean, such direct bonding could be prevented for some fraction of the dye molecules. Sample contamination is therefore also a potential source for a systematic weakening of the interfacial electronic coupling.

4. Conclusions

A series of perylene-sensitized TiO_2 nanoparticles has been studied theoretically from first principles using density functional theory calculations. Calculated electronic properties have been directly compared to experimental information about heterogeneous electron-transfer rates in the femtosecond time regime. Calculated approximate heterogeneous electron-transfer rates agree with the experimental values to within a factor of 2, and the trends for relative rates are found to agree well with the experimental results. This suggests that this kind of supramolecular calculation of dye molecules on nanoparticles can be used to predict how changes to the structure or composition of dye-sensitized semiconductor systems will affect their ultrafast electron-transfer properties. As there are no problems to accommodate larger adsorbates in the cluster approach, we believe that this approach will prove to be very useful for studies of large heterosupramolecular systems containing both organic and organometallic photo- and redox-centers attached to semiconductor nanoparticles.

Acknowledgment. The Göran Gustafsson Foundation and the Magnus Bergvall Foundation are gratefully acknowledged for financial support. We also thank the Swedish National Supercomputer Center (NSC) for generous allocations of computer resources. P.P. and M.J.L. acknowledges Prof. Sten Lunell, Uppsala University, and Dr. Lars Ojamäe, Linköping University, for stimulating discussions.

References

(1) Miller, R. J. D.; McLendon, G. L.; Nozik, A. J.; Schmickler, W.; Willig, F. *Surface Electron Transfer Processes*; VCH Publishers: 1995.

(2) Kavarnos, G. J. *Fundamentals of Photoinduced Electron Transfer*; VCH Publishers: 1993.

(3) Hagfeldt, A.; Grätzel, M. *Chem. Rev.* **1995**, *95*, 49.

(4) Hagfedt; A.; Grätzel, M. *Acc. Chem. Res.* **2000**, *33*, 269.

(5) O'Regan, B.; Grätzel, M. *Nature* **1991**, *353*, 737.

(6) Kalyanasundadaram, K.; Grätzel, M. *Coord. Chem. Rev.* **1998**, *95*, 49.

(7) Galoppini, E. *Coord. Chem. Rev.* **2004**, *248*, 1161.

(8) Schnadt, J. et al. *Nature* **2002**, *418*, 620.

(9) Zimmermann, C.; Willig, F.; Ramakrishna, S.; Burfeindt, B.; Pettinger, B.; Eichberger, R.; Storck, W. *J. Phys. Chem. B* **2001**, *105*, 9245.

(10) Benkö, G.; Kallioinen, J.; Korppi-Tommola, J. E. I.; Yartsev, A.; Sundström, V. *J. Am. Chem. Soc.* **2002**, *124*, 489.

(11) Huber, R.; Moser, J. E.; Grätzel, M.; Wachtveitl, J. *J. Phys. Chem. B* **2002**, *106*, 6494.

(12) Asbury, J. B.; Hao, E.; Wang, Y.; Lian, T. *J. Phys. Chem. B* **2001**, *105*, 4545.

(13) Marcus, R. A.; Sutin, N. *Biochim. Biophys. Acta* **1985**, *811*, 265.

(14) Lanzafame, J. M.; Palese, S.; Wang, D.; Miller, R. J. D.; Muenter, A. A. *J. Phys. Chem.* **1994**, *98*, 11020.

(15) Marcus, R. A. *J. Chem. Phys.* **1965**, *43*, 679.

(16) Willig, F.; Zimmermann, C.; Ramakrishna, S.; Storck, W. *Electrochim. Acta* **2000**, *45*, 4565.

(17) *Topics in Current Chemistry;* Gerischer, H., Willig, Boschke, F. F. L., Eds.; Springer: Berlin, 1976; Vol. 61, p 31.

(18) Newton, M. D. *Chem. Rev.* **1991**, *91*, 767.

(19) Adams, D. M. et al. *J. Phys. Chem. B* **2003**, *107*, 668.

(20) Ramakrishna, S.; Willig, F.; May, V.; Knorr, A. *J. Phys. Chem. B.* **2003**, *107*, 607.

(21) Wang, L. X.; Ernstorfer, R.; Willig, F.; May, V. *J. Phys. Chem. B* **2005**, *109*, 9589.

(22) Persson, P.; Bergström, R.; Ojamäe, L.; Lunell, S. *Adv. Quantum Chem.* **2002**, *41*, 203.

(23) Persson, P.; Stashans, A.; Bergström, R.; Lunell, S. *Int. J. Quantum Chem.* **1998**, *70*, 1055.

(24) Persson, P.; Lunell, S. *Sol. Energy Mater. Sol. Cells* **2000**, *63*, 139.

(25) Persson, P. et al. *J. Chem. Phys.* **2000**, *112*, 3945.

(26) Persson, P.; Bergström, R.; Lunell, S. *J. Phys. Chem. B* **2000**, *104*, 10348.

(27) Persson, P.; Lunell, S.; Ojamäe, L. *Chem. Phys. Lett.* **2000**, *364*, 469.

(28) Stier, W.; Prezhdo, O. V. *J. Phys. Chem. B* **2002**, *106*, 8047.

(29) Redfern, P. C.; Zapol, P.; Curtiss, L. A.; Rajh, T.; Thurnauer, M. C. *J. Phys. Chem. B* **2003**, *107*, 11419.

(30) Rego, L. G. C.; Batista, V. S. *J. Am. Chem. Soc.* **2003**, *125*, 7989.

(31) Stier, W.; Duncan, W. R.; Prezhdo, O. V. *Adv. Mater.* **2003**, *16*, 240.

(32) De Angelis, F.; Tilocca, A.; Selloni, A. *J. Am. Chem. Soc.* **2004**, *126*, 15024.

(33) Persson, P.; Lundqvist, M. J. *J. Phys. Chem. B* **2005**, *109*, 11918.

(34) Vega-Arroyo, M.; LeBreton, P. R.; Rajh, T.; Zapol, P. Curtiss, L. A. *Chem. Phys. Lett.* **2005**, *406*, 306.

(35) Rego, L. G. C.; Abuabara, S. G.; Batista, V. S. *J. Chem. Phys.* **2005**, *122*, 154709.

(36) Noguera, C *Surf. Rev. Lett.* **2001**, *8*, 121.

(37) Bredow, T.; Jug, K. *J. Phys. Chem.* **1995**, *99*, 285.

(38) Persson, P.; Gebhardt, J. C. M.; Lunell, S. *J. Phys. Chem. B* **2003**, *107*, 3336.

(39) Ernstorfer, R. Spectroscopic investigation of photoinduced heterogeneous electron transfer, Ph.D. Thesis, Freie Universität Berlin, Germany, 2004.

(40) Gundlach, L. Surface Electron-Transfer Dynamics in the Presence of Organic Chromophores, Ph.D. Thesis, Freie Universität Berlin, Germany 2005.

(41) *Gaussian 03, Revision C.02*; 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.: Wallingford, CT, 2004.

(42) Koch, W.; Holthausen, M. *A Chemist*'s *Guide to Density Functional Theory*; Wiley-VCH: Weinheim, 2001.

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

(44) Halasinski, T. M.; Weisman, J. L.; Ruiterkamp, R.; Lee, T. J.; Salama, F.; Head-Gordon, M. *J. Phys. Chem. A* **2003**, *107*, 3660.

(45) (a) te Velde, G.; Bickelhaupt, F. M.; van Gisbergen, S. J. A.; Fonseca Guerra, C.; Baerends, E. J.; Snijders, J. G.; Ziegler, T. *J. Comput. Chem.* **2001**, *22*, 931. (b) Fonseca Guerra, C.; Snijders, J. G.; te Velde, G.; Baerends, E. J. *Theor. Chem. Acc.* **1998**, *99*, 391. (c) Baerends, E. J.; Autschbach, J. A.; Bacrces, A.; Bo, C.; Boerrigter, P. M.; Cavallo, L.; Chong, D. P.; Deng, L.; Dickson, R. M.; Ellis, D. E.; Fan, L.; Fischer, T. H.; Fonseca Guerra, C.; van Gisbergen, S. J. A.; Groeneveld, J. A.; Gritsenko, O. V.; Graning, M.; Harris, F. E.; van den Hoek, P.; Jacobsen, H.; van Kessel, G.; Kootstra, F.; van Lenthe, E.; Osinga, V. P.; Patchkovskii, S.; Philipsen, P. H. T.; Post, D.; Pye, C. C.; Ravenek, W.; Ros, P.; Schipper, P. R. T.; Schreckenbach, G.; Snijders, J. G.; Sola, M.; Swart, M.; Swerhone, D.; te Velde, G.; Vernooijs, P.; Versluis, L.; Visser, O.; van Wezenbeek, E.; Wiesenekker, G.; Wolff, S. K.; Woo, T. K.; Ziegler, T. ADF2002.03, SCM, Theoretical Chemistry, Vrije Universiteit, Amsterdam, The Netherlands, http://www.scm.com.

(46) Muscat, J. P.; Newns, D. M. *Prog. Surf. Sci.* **1978**, *9*, 1.

(47) Cohen-Tannoudji, C.; Diu, B.; Laloe, F. *Quantum Mechanics;* J. Wiley and Sons: Paris, 1977; Vol. 2.

(48) Vittadini, A.; Selloni, A.; Rotzinger, F. P.; Grätzel, M. *J. Phys. Chem. B* **2000**, *104*, 1300.

(49) Diebold, U. *Surf. Sci. Reports* **2003**, *48*, 53.

CT050141X

JCTC Journal of Chemical Theory and Computation

QM/MM Free-Energy Perturbation Compared to Thermodynamic Integration and Umbrella Sampling: Application to an Enzymatic Reaction

Johannes Kästner, Hans Martin Senn, Stephan Thiel, Nikolaj Otte, and Walter Thiel*

Max-Planck-Institut für Kohlenforschung, Kaiser-Wilhelm-Platz 1, D-45470 Mülheim an der Ruhr, Germany

Received October 14, 2005

Abstract: We used the free-energy perturbation (FEP) method in quantum mechanics/molecular mechanics (QM/MM) calculations to compute the free-energy profile of the hydroxylation reaction in the enzyme *p*-hydroxybenzoate hydroxylase (PHBH). *k* statistics were employed to analyze the FEP sampling including estimation of the sampling error. Various approximations of the free-energy perturbation method were tested. We find that it is adequate not only to freeze the density of the QM part during the dynamics at frozen QM geometry but also to approximate this density by electrostatic-potential-fitted point charges. It is advisable to include all atoms of a QM/MM link in the perturbation. The results of QM/MM-FEP for PHBH are in good agreement with those of thermodynamic integration and umbrella sampling.

I. Introduction

The free energy is the measure for the driving force of a chemical reaction. It can be calculated by a variety of methods. As bonds are broken and formed in chemical reactions, quantum mechanical methods are required for the calculation of free-energy differences. Standard electronic-structure methods provide the internal energy ΔU at zero temperature, $T = 0$ K. The Helmholtz free energy, $\Delta A = \Delta U - T\Delta S$, at a finite temperature also includes the entropy change ΔS. While an approximation for ΔS may be obtained from the harmonic frequencies of the system, it can be more accurately computed by sampling along a reaction coordinate.

In principle, the free-energy change along a reaction coordinate ξ may be calculated directly from the distribution function of ξ obtained from a molecular-dynamics (MD) simulation. Such a calculation may be accelerated by umbrella sampling (US).[1,2] This method applies a restraint (bias) to the reaction coordinate. In the limit of an infinitely strong bias, that is, a constraint, the method becomes equivalent[3] to thermodynamic integration (TDI).[4−7] The free-energy change may then be determined by integration of the mean force on this constraint. Both methods, thermodynamic integration and umbrella sampling, are based on an exhaustive sampling of the phase space. For large molecules, this is currently still impractical when using computationally demanding ab initio or density functional methods because of the prohibitive computational effort. This holds true even if such methods serve as QM components in QM/MM approaches where the reactive center is described by quantum mechanics (QM) and the environment by molecular mechanics (MM).

The sampling problem has been addressed by number of different approaches.[8−17] Here, we focus on a QM/MM-FEP treatment[8] that applies the free-energy perturbation (FEP) method[18] to QM/MM simulations (see ref 8 for a comparison with previously available approaches[9,11−14]). In QM/MM-FEP, only the computationally less demanding MM part is sampled, while the demanding QM part is kept frozen. Free-energy perturbation used in this manner includes the following approximations: (1) The entropy change within the QM part is not sampled but estimated from the harmonic approximation. (2) Commonly,[8,16,19−22] the density of the fixed QM system is not only frozen but approximated by electrostatic potential (ESP) charges when calculating its interaction with the MM part.

We have tested these approximations on a biological system, the enzyme *p*-hydroxybenzoate hydroxylase[23] (PHBH; EC 1.14.13.2). While natively catalyzing the transformation

* Corresponding author. E-mail: thiel@mpi-muelheim.mpg.de.

10.1021/ct050252w CCC: $33.50
Published on Web 02/15/2006

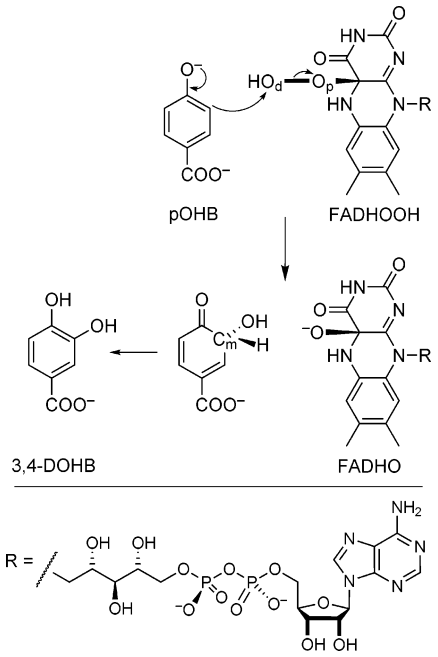


Figure 1. Schematic view of the rate-determining OH-transfer reaction catalyzed by the enzyme PHBH.

of *p*-hydroxybenzoate (pOHB) to 3,4-dihydroxybenzoate (3,4-DOHB), PHBH has also been proposed as a biocatalyst for the hydroxylation of halogenated pOHB derivatives.[24] During the catalytic cycle, the flavine cofactor (FAD: flavine−adenine dinucleotide) is reduced to $FADH_2$ by NADPH (nicotinamide−adenine dinucleotide phosphate, reduced form). It then reacts with molecular oxygen to form the flavin hydroperoxide (FADHOOH), shown in Figure 1. In what is believed to be the rate-determining step,[25−27] FADHOOH hydroxylates the substrate pOHB, yielding FADHO and a hydroxycyclohexadienone that tautomerizes rapidly to the aromatic 3,4-DOHB. FADHO is finally protonated to FADHOH, loses water, and regenerates the oxidized form FAD.

The rate-determining step is shown in Figure 1. The substrate, in its dianionic form,[25,26] is hydroxylated in an aromatic electrophilic substitution reaction. From temperature-dependent measurements of the overall rate, the activation energy was estimated as 49 kJ mol^{-1} at pH 8.0.[28] AM1 has been shown to overestimate the reaction barrier[29,30] but to yield fairly accurate structures, except for the underestimated peroxide O−O bond length.[29,31]

While the QM/MM-FEP method itself is not new,[8] the purpose of this work is to validate its approximations by calculating their influence on the resulting free-energy difference. We carefully estimate the effects of different treatments of the electrostatic QM/MM interaction. Problems occurring in the perturbation of link atoms and their solutions are also discussed. Moreover, we point out how to use *k* statistics to analyze the FEP results. The QM/MM-FEP method is designed to be applicable to QM/MM setups with demanding QM methods. However, to test the method, we employed the fast semiempirical AM1 Hamiltonian.[32] The methodological issues raised by QM/MM-FEP are expected to be similar for AM1 and higher-level QM methods, and the choice of AM1 allows us to investigate these issues efficiently and to assess various QM/MM-FEP approximations by comparisons against full QM calculations. Thus, the primary aim of this study is to test QM/MM-FEP on a real-world example, rather than to accurately reproduce the experimental activation barrier.

II. Methods

II.A. QM/MM Free-Energy Perturbation. In its original formulation,[18] free-energy perturbation is defined via an unperturbed Hamiltonian and a perturbation term ΔE_{pert}. Sampling ΔE_{pert} makes it possible to calculate the free-energy difference of the perturbation by exponential averaging: $\Delta A = -1/\beta \ln\langle\exp(-\beta\Delta E_{pert})\rangle$. $\langle x\rangle$ denotes a canonical average, $\beta = (k_B T)^{-1}$, and k_B is the Boltzmann constant. The term "perturbation" is somewhat misleading since the theory is exact and does not correspond to a perturbation ansatz in the usual sense. Applied to QM/MM simulations,[8] the unperturbed Hamiltonian corresponds to the QM/MM energy expression of a system where the QM atoms are fixed. The perturbation corresponds to a geometry step of the QM atoms. The phase space sampled is restricted to the degrees of freedom of the MM atoms. This allows one to calculate a free-energy profile in QM/MM simulations using demanding QM methods.

We briefly restate the formalism of QM/MM-FEP to point out the approximations used. The total energy of a QM/MM calculation may be written as

$$E_{total} = E_{qm}(\mathbf{r}_{qm}) + E_{qm/mm}(\mathbf{r}_{qm},\mathbf{r}_{mm}) + E_{mm}(\mathbf{r}_{mm}) \quad (1)$$

with E_{qm} depending on the coordinates of the QM atoms $\mathbf{r}_{qm}$, and E_{mm} depending on the coordinates of the MM atoms $\mathbf{r}_{mm}$. The term $E_{qm/mm}(\mathbf{r}_{qm},\mathbf{r}_{mm})$ includes all energy contributions coupling the QM and the MM parts:

$$E_{qm/mm} = E_{vdW} + E_Q + E_{FF} \quad (2)$$

that is, the van der Waals interaction E_{vdW}, the electrostatic interaction E_Q, and the force field terms E_{FF} of the junctions. The latter come from covalent bonds between QM and MM atoms. Within the electrostatic embedding scheme, the MM point charges polarize the QM part, and the electrostatic interaction between the QM and MM parts, E_Q, is therefore included in the energy provided by the QM code. Under the convention of eq 2, E_Q thus has to be calculated separately for obtaining $E_{qm/mm}$. From the QM electronic energy in the point-charge field, $\langle\Psi|\mathscr{H}|\Psi\rangle$, one obtains $E_{qm} = \langle\Psi|\mathscr{H}|\Psi\rangle - E_Q$.

We will divide our discussion of QM/MM-FEP into three steps: (1) calculation of an energy profile of the reaction using constrained optimizations, (2) calculation of the energy of the perturbation, ΔE_{pert}, and (3) sampling of ΔE_{pert}. While this separation is conceptually sensible, steps 2 and 3 are coupled in practical simulations.

II.A.1. Optimization. A reaction coordinate $\xi(\mathbf{r}_{qm})$ depending only on QM positions is defined. The reaction is split into discrete windows, each characterized by a value ξ_i. Constraining the reaction coordinate to some ξ_i, all other QM and MM degrees of freedom are optimized for each window *i*. This results in a set of minimum-energy geom-

etries and a profile of the internal energy of the reaction at zero temperature.

II.A.2. Perturbation. The optimized structures serve as unperturbed and perturbed structures in turn: When structure i is perturbed with structure $i + 1$, the energy of perturbation is given by

$$\Delta E_{\text{pert}}^{i\to i+1} = \underbrace{E_{\text{qm/mm}}(\mathbf{r}_{\text{qm}}^{i+1}, \mathbf{r}_{\text{mm}}^{i})}_{\text{perturbed}} - \underbrace{E_{\text{qm/mm}}(\mathbf{r}_{\text{qm}}^{i}, \mathbf{r}_{\text{mm}}^{i})}_{\text{unperturbed}} \quad (3)$$

Thus, the unperturbed energy is calculated with all atoms at their positions of the optimized window i. The perturbed energy is calculated with the MM positions of window i and the QM positions of window $i + 1$. Equation 3 defines the "forward perturbation". In the "backward perturbation", window $i + 1$ is perturbed with window i:

$$\Delta E_{\text{pert}}^{i+1\to i} = E_{\text{qm/mm}}(\mathbf{r}_{\text{qm}}^{i}, \mathbf{r}_{\text{mm}}^{i+1}) - E_{\text{qm/mm}}(\mathbf{r}_{\text{qm}}^{i+1}, \mathbf{r}_{\text{mm}}^{i+1}) \quad (4)$$

II.A.3. Sampling. The free-energy change between window i and window $i + 1$ is given by

$$\Delta A^{i\to i+1} \approx \Delta E_{\text{qm}}^{i\to i+1} + \Delta A_{\text{qm/mm}}^{i\to i+1} \quad (5)$$

$\Delta E_{\text{qm}}^{i\to i+1}$ is the difference of the QM energies between the windows i and $i + 1$. $\Delta A_{\text{qm/mm}}^{i\to i+1}$ incorporates the change in the free energy due to the QM/MM interactions as well as the MM part. It is obtained from sampling

$$\Delta A_{\text{qm/mm}}^{i\to i+1} = -\frac{1}{\beta} \ln\langle \exp(-\beta \Delta E_{\text{pert}}^{i\to i+1}) \rangle_{\text{mm},i} \quad (6)$$

The energy difference is sampled at window i, meaning that the MM atoms move according to the forces from the QM part of window i. The average is only taken over the MM coordinates, since the QM coordinates are always frozen in the MD simulations. Forward ($i \to i + 1$) and backward ($i + 1 \to i$) perturbation converge to the same energy difference, with opposite sign. As $\Delta E_{\text{qm/mm}}^{i\to i+1} + \Delta E_{\text{mm}}^{i\to i+1}$ is known from the reaction profile, the knowledge of $\Delta A_{\text{qm/mm}}^{i\to i+1}$ allows an estimate of the corresponding entropic contributions.

Values for $\Delta A_{\text{qm/mm}}^{i\to i+1}$ of a typical simulation, including the error bar defined by eq 17, are shown in Figure 2. Summation of the results of eq 5 provides $\Delta A(\xi)$ on a grid provided by the ξ_i values of the different windows. Minima and maxima are determined by interpolating three consecutive values of $\Delta A(\xi)$ with a second-order polynomial.

II.B. Entropic Effects of the QM Part. In eq 5, the entropy and finite-temperature effects in the QM part have been neglected. These can be taken into account by calculating the harmonic frequencies of the QM part and applying standard methods from statistical thermodynamics to evaluate the difference $\Delta A_{\text{qm}} - \Delta E_{\text{qm}}$ for the stationary points of interest (minima, transition states).[8,33]

II.C. Electrostatic Interaction in QM/MM-FEP. During the MD sampling, the structure of the QM part is kept frozen at the optimized geometry of either window i or $i + 1$ for the forward and backward perturbation, respectively. Instead of calculating E_{Q} from a density obtained from full self-consistent field (SCF) iterations in each MD step, one may

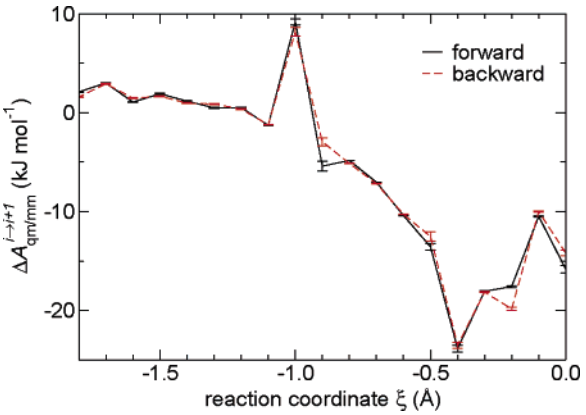


Figure 2. $\Delta A_{\text{qm/mm}}^{i\to i+1}$ obtained from forward perturbation and backward perturbation including the error bar. Differences between forward and backward perturbations are mainly caused by incomplete sampling. The spike at $\xi = -1.0$ Å is caused by the perturbation between two manifolds, see section III.A.

introduce the approximation to freeze the density ρ, that is, to neglect changes in the polarization of the density caused by the varying MM coordinates during the MD run for a given window. Calculating E_{Q} from a fixed density requires evaluating one-electron integrals only, but no SCF iterations, and is thus computationally less expensive.

In a further approximation, charges which reproduce the electrostatic potential (ESP charges) are commonly used instead of the full density to calculate E_{Q}.[8,16,19−22] As the point charges should reproduce the energy and forces generated by the full density, ESP charges are well-suited. This allows one to completely avoid QM calculations in the sampling runs. When an accurate—but slow—QM method is used, the vast majority of the calculation time is, therefore, spent on obtaining the optimized structures.

We tested both approximations, fixed ρ and ESP charges, by comparison to the full SCF density using the fast AM1 method. In section III.C, we show that both approximations are well-justified, at least for the system under investigation.

II.D. Link Atoms and Their Perturbation. In our QM/MM setup,[34] covalent bonds between the QM and the MM parts, so-called junction bonds, are capped by link atoms. These are only treated by the QM code and are invisible to the MM code. The link atom is placed on the line connecting the QM atom and the MM atom of the junction bond at a constant distance from the former. Forces on the link atom are remapped to the QM and MM junction atoms. The stretching of the junction bond is described at the MM level. As the link atom is intrinsically of QM nature, its position is constrained in the MD sampling. We also freeze the MM atom of the junction.

To calculate $E_{\text{qm/mm}}(\mathbf{r}_{\text{qm}}^{i+1}, \mathbf{r}_{\text{mm}}^{i})$ in eq 3, the QM geometry is taken from window $i + 1$, while the MM geometry is the one from window i. The question of what to do with the link atoms arises. There are four possibilities to treat the position of the junction atoms, illustrated in Figure 3: (1) Only the QM atom is moved to its position in window $i + 1$; the link and the MM atom remain at their positions in window i. (2) The link atom is placed on the line connecting

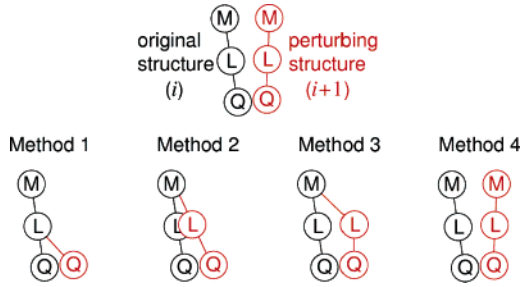


Figure 3. Four methods of perturbing link atoms. Q, L, and M refer to the QM atom, the link atom, and the MM atom of the junction, respectively.

the QM atom of window $i + 1$ and the MM atom of window i. The MM atom remains at its position in window i. (3) QM and link atoms are moved to their positions in window $i + 1$; the MM atom remains at i. (4) All three atoms are moved to their positions in window $i + 1$.

The drawback of methods 1 and 2 is that the link atom is moved from the position where its density has been calculated. In methods 1 and 3, the position of the link atom is inconsistent as it does not lie anymore on the line connecting the QM and MM junction atoms. Method 4 thus emerges as the most consistent and promising choice, although it involves the largest perturbation as all three atoms are moved.

II.E. System Setup. The enzyme PHBH was treated in analogy to our previous work.[29,35] The initial geometry was based on a crystal structure (see PDB file 1IUW).[36] The enzyme, consisting of 394 amino acids, 219 crystallographic water molecules, the FADHOOH cofactor, and fully deprotonated *p*-hydroxybenzoate, was solvated in a cubic water box. After a series of structure optimizations and MD runs, a production run was performed under periodic boundary conditions in the canonical (*NVT*) ensemble at $T = 300$ K with restraints acting on the cofactor and the substrate. A snapshot after 40 ps was used as the starting point for this study. All water molecules further than 11 Å from any protein atom were discarded. The water molecules in the outer solvation shell between 2.9 and 11 Å were kept rigid, and all atoms inside were left free. This resulted in 6245 protein atoms, 102 atoms of the cofactor and the substrate, and 2445 water atoms, all of which were free to move. In total, there are 22 772 atoms (8792 free and 13 980 fixed in the outer solvation shell).

The QM/MM setup was chosen as follows. The QM region consisted of 49 atoms: the substrate *p*-hydroxybenzoate (in its dianionic form) and the isoalloxazine part of the cofactor FADHOOH. There is one covalent bond between the QM and the MM parts, which was saturated by a H link atom. The QM part was described with the semiempirical AM1 Hamiltonian[32] and the protein environment with the GROMOS force field. We used the QM/MM approach as implemented in the ChemShell software package.[34] ChemShell provided the optimizer, the MD driver, and the interfaces to the MNDO99[37] and GROMOS96[38] codes. The ability to perform QM/MM-FEP calculations was implemented into ChemShell. The electrostatic interaction between the QM and the MM atoms was treated by including all MM point charges in the QM Hamiltonian. The charge-shift scheme[34] was applied at the junction. The MM-MM electrostatic interactions were evaluated explicitly for all atom pairs with a distance of up to 14 Å and approximated by a generalized Poisson−Boltzmann reaction field[39] with $\epsilon_r = 54.0$ beyond. The SCF convergence criterion was 10^{-8} eV.

The MD snapshot (see above) of the reactant was the starting point for structure optimizations. The whole system, that is, all atoms except the frozen outer solvation shell, was optimized in hybrid delocalized internal coordinates[40,41] (HDLC) to a convergence criterion for the maximum gradient component of $0.45 \times 10^{-3}\ E_h\ a_0^{-1}$, using a limited-memory quasi-Newton algorithm[42] (L-BFGS; BFGS: Broyden−Fletcher−Goldfarb−Shanno). This led to the reactant state. For the following transition-state search and optimizations, we defined an active region of about 3300 atoms, composed of the QM part and all residues with at least one atom within 15 Å of the substrate.

For the transition-state search, we chose a reaction core of nine atoms directly involved in the OH transfer. The OH group was manually displaced toward the expected transition state. The microiterative transition-state search in HDLCs proceeded as follows: L-BFGS steps were performed for all atoms of the active region except the reaction core until the environment was converged to within a maximum gradient component of $0.45 \times 10^{-3}\ E_h\ a_0^{-1}$. One partitioned rational function optimizer[43] (P-RFO) step was then performed for the reaction core, using an explicit Hessian. Continuing with the L-BFGS optimization, this process was iterated until the largest gradient component in the core was less than $1.35 \times 10^{-3}\ E_h\ a_0^{-1}$.

The difference of two bond lengths was defined as the reaction coordinate:

$$\xi = d(O_d{-}O_p) - d(C_m{-}O_d) \qquad (7)$$

See Figure 1 for atom labeling. Note that the sign of ξ is opposite in ref 35. The transition state is at $\xi(\mathrm{TS}) = -0.41$ Å. A reaction profile in intervals of 0.1 Å was calculated to define the windows for the FEP simulations. The starting point was $\xi = -0.4$ Å, which was generated from the TS geometry by moving O_d by 0.01 Å along the reaction coordinate. In each of the windows, the reaction coordinate was constrained, and all other degrees of freedom in the active region were optimized. The profile ranged from $\xi = -1.8$ to $+1.8$ Å. At these optimized structures, ESP charges were calculated by fitting the potential at the positions of the 200 MM atoms nearest to the QM atoms.

The MD simulations were started from these optimized structures. The QM part, the first MM atom of the junction, and the outer solvation shell were frozen. All other 8742 atoms were equilibrated for 30 ps for the window at $\xi = -0.4$ Å. During the heating phase, the first 10 ps of this equilibration, a Berendsen thermostat[44] was used. In all other MD simulations, a canonical (*NVT*) ensemble at $T = 300$ K was generated by a Nosé−Hoover chain thermostat[45−48] with a chain length of 4 and a characteristic period of 20 fs, corresponding to a thermostat wavenumber of 375 cm^{-1}. Newton's equations of motion were integrated with a reversible noniterative leapfrog-type integrator[49] with a time

step of 1 fs. To ensure energy conservation at this time step, all hydrogen atoms were assigned the mass of deuterium and the free water molecules were kept internally rigid using SHAKE constraints.[50] To prepare the subsequent window, the QM part of the equilibrated window with $\xi = -0.4$ Å was replaced by the QM part with $\xi = -0.3$ Å. This system was again equilibrated for 10 ps. In this manner, all windows were equilibrated consecutively.

The FEP production runs were performed for the forward and the backward perturbation for 10 ps in each window, unless noted otherwise. Equilibration of the system with respect to ΔE_{pert} was tested as described elsewhere[51] by testing for the lack of a trend in the coarse-grained average and its variance, for normality, and for a lack of correlation. Whenever the tests for trend showed that ΔE_{pert} or its variance were not stationary over the whole range of the production run, MD steps from the beginning of the simulation were dropped until stationarity was reached. This was done separately for the forward and the backward perturbation data of each window. The range with stationary ΔE_{pert} was then used for the analysis.

To compare the results of FEP with other free-energy methods, we performed TDI and US simulations. Methodology and detailed results of TDI simulations have been reported previously.[35] The reaction coordinate was constrained in intervals of 0.1 Å. The force of constraint was sampled and integrated along ξ to compute the free energy. In the simulations for umbrella sampling, the constraint was replaced by a harmonic restraint of the form $w_i = K/2(\xi - \xi_i)^2$ with $K = 0.18\ E_h\ a_0^{-2}$. The values of ξ_i were chosen to be the same as in the TDI simulations. The structures resulting from the TDI sampling were used as starting structures for the US simulations. After a re-equilibration of 2 ps, which is necessary because the constraints were replaced by restraints, the system was sampled until the mean and the variance of ξ were trend-free according to the Mann−Kendall test[51,52] over at least 8 ps. The data of these 8 ps were then used for the analysis. We used the weighted histogram analysis method (WHAM)[53,54] as well as umbrella integration[3] to combine the different windows of the umbrella sampling simulations.

To test the approximation of the full QM density by ESP charges, we used not only AM1 but also density functional theory (DFT).[55,56] These calculations were done with the TURBOMOLE[57−61] code (version 5.7.1) interfaced to ChemShell, using the BP86 functional.[62−66] The DFT optimization was started from the AM1 geometry and carried out with the TZVP[67] basis set. At the optimized geometry, the self-consistent electron density was computed with the aug-cc-pVTZ[68,69] basis, which includes polarization and diffuse functions. ESP charges were fitted to this density to compare the forces on the MM atoms obtained from the density and the ESP charges.

II.F. Statistical Analysis of FEP Results. In free-energy perturbation, exponential averages of the form $\langle \exp(-\beta \Delta E_{\text{pert}}) \rangle$ have to be evaluated.[70] These are dominated by small values of ΔE_{pert}, which are poorly sampled. Thus, the result may strongly depend on the random occurrence of low values of ΔE_{pert} in the trajectory.

It is more efficient to use an expansion of $\langle \exp(-\beta \Delta E_{\text{pert}}) \rangle$ rather than the direct exponential average. $\langle \exp(-\beta \Delta E_{\text{pert}}) \rangle$ can be expressed as a cumulant expansion[18,71−74]

$$\ln\langle e^{-\beta \Delta E_{\text{pert}}} \rangle = \sum_{i=1}^{\infty} \frac{(-\beta)^i}{i!} \kappa_i \tag{8}$$

with the κ_i being the cumulants. They depend on the first and higher moments of the distribution of ΔE_{pert}.

In the case of an MD simulation, the distribution itself is not available; thus, the cumulants cannot be calculated directly. Only the sample values drawn from the distribution are available. The ith k statistic k_i is the unique symmetric unbiased estimator of the cumulant κ_i.[71,74] The first four k statistics are given[71,74] in terms of the sample size N, the sample mean $\langle \Delta E_{\text{pert}} \rangle$, and the sample central moments m_i, with $m_i = 1/N \sum_{j=1}^{N} (\Delta E_{\text{pert},j} - \langle \Delta E_{\text{pert}} \rangle)^i$:

$$k_1 = \langle \Delta E_{\text{pert}} \rangle \tag{9}$$

$$k_2 = \frac{N}{N-1} m_2 \tag{10}$$

$$k_3 = \frac{N^2}{(N-1)(N-2)} m_3 \tag{11}$$

$$k_4 = \frac{N^2[(N+1)m_4 - 3(N-1){m_2}^2]}{(N-1)(N-2)(N-3)} \tag{12}$$

For a normally distributed sample, all cumulants κ_i for $i \geq 3$ vanish. In general, the distribution of ΔE_{pert} taken from an MD simulation in a given window is very close to a normal distribution. (A special case where this is not true will be discussed in section III.C.) We, therefore, truncate the cumulant expansion after the second term. Only the mean and the variance of ΔE_{pert} are then required to calculate the estimate of $\langle \exp(-\beta \Delta E_{\text{pert}}) \rangle$. These, however, depend much less on the equilibration of the system than the direct average of $\exp(-\beta \Delta E_{\text{pert}})$ or the higher cumulants, which are more strongly influenced by rarely occurring small values of ΔE_{pert}.

To calculate an error bar for ΔA, we use the estimators of the variances of the k statistics:[71,74]

$$\widehat{\text{var}}(k_1) = \frac{k_2}{N} \tag{13}$$

$$\widehat{\text{var}}(k_2) = \frac{2N{k_2}^2 + (N-1)k_4}{N(N+1)} \tag{14}$$

In the special case of a normal parent distribution, the estimator of $\text{var}(k_3)$ is

$$\widehat{\text{var}}(k_3) = \frac{6N(N-1){k_2}^3}{(N-2)(N+1)(N+3)} \tag{15}$$

Thus, we use $-1/\beta \ln\langle \exp(-\beta \Delta E_{\text{pert}}) \rangle = \overline{\Delta A} \pm 2s$ as the confidence interval, approximating the Student t fractile at a confidence level of 95% by 2 for large N and estimating s^2 from error propagation.

$$\overline{\Delta A} = \langle \Delta E_{\text{pert}} \rangle - \frac{\beta}{2} k_2 \tag{16}$$

$$s^2 = \widehat{\text{var}}(k_1) + \frac{\beta^2}{4}\widehat{\text{var}}(k_2) + \frac{\beta^4}{6^2}\widehat{\text{var}}(k_3) \tag{17}$$

Note that this error measure only accounts for the statistical fluctuations of the MD run. It includes neither errors caused by incomplete sampling nor errors caused by the method itself, such as the choice of the QM region or the intrinsic accuracy of the QM or the MM method.

III. Results and Discussion

III.A. Structural Issues. In the reaction, the transfer of an OH group from the hydroperoxy group to *p*-hydroxybenzoate takes place. The stationary points that emerged from the structure optimizations are shown in Figure 4. In the transition state, the hydrogen atom of the OH group is stabilized by a hydrogen bond to the backbone amide oxygen of Pro293. During the stepwise structural optimizations along the reaction coordinate, this hydrogen bond remained present. In the case of the product state, a local minimum was found around $\xi = 1.0$ Å, see Figure 5. The hydrogen bond broke at the optimization for $\xi = 1.7$ Å with a distinct energy lowering. Stepwise backward optimization resulted in a second minimum around $\xi = 1.5$ Å, which is 8.5 kJ mol^{-1} lower than the first local minimum. In this case, a different hydrogen bond, OH···O$_p$, was formed. Further backward optimization increased the energy above the curve obtained with the OH···O(Pro293) hydrogen bond. The system thus shows a hysteresis. Note that all values given here refer to energies and reaction coordinates on the reaction profile and were obtained from calculations with constrained ξ values in intervals of 0.1 Å. An analogous behavior was found for the reactant state. In this case, H of the OOH group does not participate in any hydrogen bond in the more stable minimum.

The occurrence of such a hysteresis shows that the chosen reaction coordinate does not account for all structural changes during the reaction. The hydrogen bonds change but do not contribute to the reaction coordinate. Introducing the term "manifold" for the set of geometries with a given hydrogen-bond pattern, the transition state belongs to a different manifold [with an OH···O(Pro293) hydrogen bond] than the product (with OH···O$_p$) and the reactant state (no such hydrogen bond involving OH). In the FEP simulations, a change between these manifolds has to be accomplished.

In the case of the reactant state, the structural changes between the two manifolds are small, and FEP calculations between them remain possible. The change from the manifold with OH···O(Pro293) to the one of the reactant state occurs in the perturbation between $\xi_i = -0.9$ Å and $\xi_{i+1} = -1.0$ Å. This results in a spike in $\Delta A^{i\to i+1}_{\text{qm/mm}}$ at $\xi = -1.0$ Å, see Figure 2, which however does not lead to a noticeable discontinuity in $A(\xi)$, shown in Figure 6, due to compensating changes in $\Delta E^{i\to i+1}_{\text{qm/mm}}$, see eq 5.

Near the product state, the structural changes between the two manifolds are too large to be overcome in a single perturbation simulation, causing a high statistical error, as

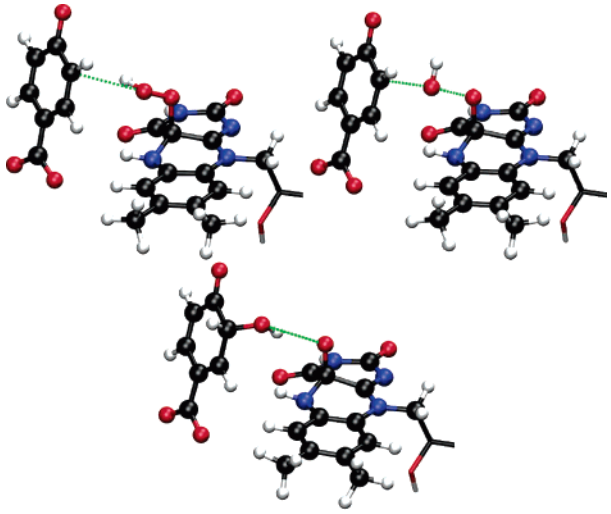


Figure 4. Reactant state, transition state, and product state of the OH-transfer reaction of PHBH. The substrate and the truncated cofactor are shown. The reaction coordinate is indicated by a dotted green line. The atoms included in the QM part are drawn as a ball-and-stick model.

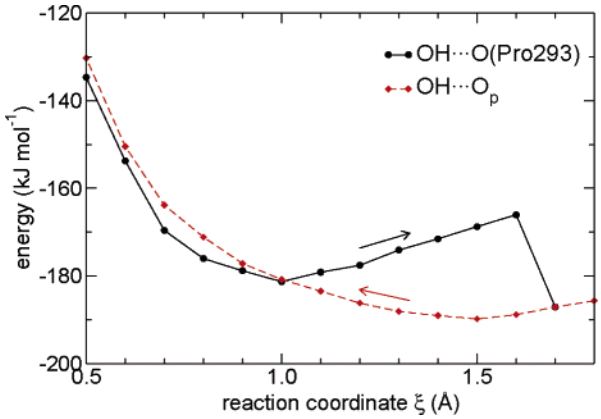


Figure 5. The two manifolds near the product state (see text). The one with a hydrogen bond OH···O(Pro293) leads to the transition state. The arrows indicate the direction in which the energy profile was calculated.

obtained from eq 17. Therefore, we used 10 intermediate windows with structures obtained from linear interpolation between the windows with $\xi_i = 0.6$ Å and $\xi_{i+1} = 0.7$ Å. With this choice, the perturbation runs converged.

The reaction path between the reactant state and the transition state will be used in section III.C to discuss the approximations used in QM/MM-FEP. First, however, we compare different methods of free-energy sampling.

III.B. Comparison of Methods. We calculated the energy profile of the complete reaction with several methods. In the FEP calculations, the link atoms were perturbed with method 4, and the sampling was done with the full, but frozen, QM density.

Figure 6 shows the comparison between the energy profiles obtained by optimization, FEP, TDI, and US. The curves have been shifted in energy to match best in the reactant and the transition state. The stationary points obtained by each method are marked, and the corresponding numerical results are given in Tables 1−3. With the use of the harmonic

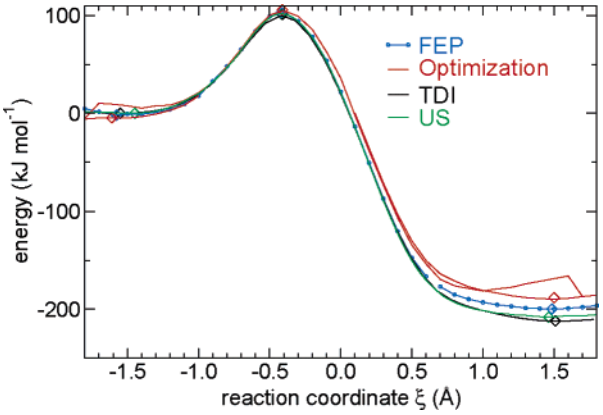


Figure 6. Energy profiles with different methods. Blue, FEP; red, optimization; black, TDI; and green, US. The two optimization curves represent two local minima near the product and the reactant (with different H bonds, see text). The gap in the FEP graph at $\xi = 0.6-0.7$ Å corresponds to the change between these two manifolds in the FEP calculations.

Table 1. Free Energies of Activation ($\Delta^{\ddagger}A$) and Reaction ($\Delta_r A$) in kJ mol^{-1} Obtained with Different Methods[a]

method	$\Delta^{\ddagger}A$	$\Delta_r A$
optimization	112.3	−184.3
FEP	108.2 ± 1.0	−198.6 ± 1.3
TDI[35]	101 ± 2	−212 ± 2
US	101.5	−208.1

[a] Thermal and entropic corrections for the QM region are included in the values given for FEP and optimization (harmonic approximation).

Table 2. Contributions to the Free-Energy Changes ΔA in kJ mol^{-1}[a]

	forward barrier	reaction energy
ΔA	108.2	−198.6
ΔE_{qm}	80.2	−350.3
$\Delta A_{qm/mm}$	26.1	152.4
$\Delta A_{qm} - \Delta E_{qm}$	1.9	−0.7
ΔE_{mm}	−10.8	−12.7
$\Delta A_{qm/mm} - \Delta E_{qm/mm} - \Delta E_{mm}$	−4.1	−14.3
ΔE_{qm}(ZPE)	5.8	1.1

[a] $\Delta A_{qm/mm}$ was calculated by FEP; ΔE_{qm}, $\Delta E_{qm/mm}$, and ΔE_{mm} were calculated by optimization; and $\Delta A_{qm} - \Delta E_{qm}$ as well as ΔE_{qm}(ZPE) were calculated from the harmonic frequencies of the QM part at the stationary points.

Table 3. Reaction Coordinate at the Reactant State, the Transition State, and the Product State (Å) Obtained with Different Methods

method	ξ(RS)	ξ(TS)	ξ(PS)
optimization	−1.61	−0.41	1.50
FEP	−1.58	−0.42	1.49
TDI[35]	−1.55	−0.41	1.51
US	−1.45	−0.42	1.46

approximation, the thermal and entropic contributions of the QM region have been included in the values for optimization and FEP in Table 1 (but not in Figure 6).

The TDI and US approaches are expected to yield the same free-energy changes because they both sample the entire system. The differences between TDI and US results are indeed very small (Table 1) and are most probably caused by incomplete sampling. US was analyzed by umbrella integration.[3] WHAM analysis of the umbrella sampling data leads to a range of 100.1−102.3 kJ mol^{-1} for the activation barrier and −205.7 to −209.4 kJ mol^{-1} for the reaction energy, depending on the number of bins used for the analysis.

In the FEP approach, the thermal and entropic contributions to the free energy are evaluated in harmonic approximation for the QM region ($\Delta A_{qm} - \Delta E_{qm}$) and are determined by sampling the environment ($\Delta A_{qm/mm} - \Delta E_{qm/mm} - E_{mm}$). It can be seen from Table 1 that the FEP results are close to the TDI and US reference values, which supports the validity of the FEP approximations for the QM and MM regions (see above). It is obvious from Table 1 that FEP accounts for some but not all of the differences between optimization on one hand and TDI or US on the other hand.

Table 2 lists the individual contributions to the FEP free-energy changes ΔA, which are the sums of the three following energies (lines 2−4). The contribution ΔE_{qm} from the QM energy is dominant, and the term $\Delta A_{qm/mm}$ is also substantial. The thermal and entropic contributions from the QM region ($\Delta A_{qm} - \Delta E_{qm}$) are small compared to those from the environment ($\Delta A_{qm/mm} - \Delta E_{qm/mm} - \Delta E_{mm}$), emphasizing the importance of including the latter (as done in FEP). These QM/MM thermal and entropic contributions lower the barrier by 4.1 kJ mol^{-1} and make the reaction more exergonic by 14.3 kJ mol^{-1} (see Table 2), which may be related to the changes in the hydrogen bond network during the reaction.[35] Finally, it should be pointed out that zero-point vibrational corrections have not been applied to the results given (Table 1). Such corrections are not available for TDI or US but can be deduced for FEP at least for the QM region from the computed QM Hessian: the corresponding ΔE_{qm}(ZPE) values are fairly small in the present case (Table 2).

The values of the reaction coordinate are listed in Table 3. The curvature of the energy surface around the transition state is rather high; thus, ξ(TS) is nearly independent of the method. In the reactant and product states, the energy surface is flat because the reaction coordinate contains a distance which does not correspond to a chemical bond. This causes larger variations in the reaction coordinate for the minima.

III.C. Test of Approximations. *III.C.1. Frozen Density.* The use of free-energy perturbation as implemented here requires freezing the QM geometry during the perturbation sampling. We assume that it is adequate to neglect polarization of the QM density in response to the moving MM atoms in the environment. To assess this assumption, we also sampled the system with full SCF iterations. With the frozen density, we obtain a free-energy barrier $\Delta^{\ddagger}A = 106.3 \pm 0.99$ kJ mol^{-1}, while we obtain $\Delta^{\ddagger}A = 103.8 \pm 0.99$ kJ mol^{-1} for full SCF iterations, see Table 4. The difference is not negligible but would seem to be tolerable in practice, especially in view of the computational savings: SCF iterations are avoided and only one-electron integrals have to be evaluated.

Table 4. Reaction Coordinate at the Reactant State and the Transition State (Å) and Forward Free-Energy Barrier (kJ mol^{-1}) for Different Approximations Used in FEP

density	link method	ξ(RS)	ξ(TS)	$\Delta^{\ddagger}A$
frozen	4	−1.578	−0.414	106.3 ± 0.99
full SCF	4	−1.501	−0.423	103.8 ± 0.99
ESP	4	−1.564	−0.431	105.7 ± 1.18
frozen	1	−1.580	−0.416	106.5 ± 0.82
frozen	2	−1.577	−0.415	105.3 ± 0.82
frozen	3	−1.576	−0.415	105.0 ± 0.82
frozen	4	−1.578	−0.414	106.3 ± 0.99

III.C.2. Density Replaced by Point Charges. For a frozen density, the computational demands can be further reduced if $E_Q(\mathbf{r}_{qm},\mathbf{r}_{mm})$ is calculated from point charges approximating the density and not from the full density itself. This reduces the sampling to a pure MM simulation from the computational point of view and, thus, significantly reduces the computation effort. Re-equilibration of the system previously sampled with the full density was necessary after switching to point charges, as this changes the gradient of the MM atoms near the QM region. We re-equilibrated each window for 20 ps and also used production runs of 20 ps. The energies obtained with ESP charges are very close to those obtained with the full density. We obtain $\Delta^{\ddagger}A = 106.3 \pm 0.99$ kJ mol^{-1} with the full density and 105.7 ± 1.18 kJ mol^{-1} when calculating all electrostatic interactions from ESP charges, see Table 4.

In an alternative approach, one might consider calculating ΔE_{pert} from the full frozen density and the forces for the dynamics from ESP charges. Testing this approach is sensible as the evaluation of the energy is much less demanding than the evaluation of the gradient. However, numerical problems occur that can be rationalized by consideration of the theoretical basis of FEP: The exponential average of ΔE_{pert} only yields the free-energy difference when sampled over the canonical ensemble.[70] If the dynamics, that is, the averaging, is performed with ESP charges while ΔE_{pert} is still calculated from the full frozen density, the ensemble does not match with ΔE_{pert}. Thus, it is not advisable to use different expressions of E_Q for the dynamics and the perturbation.

The ESP charges are generally able to reproduce the multipoles of the QM density very well. Thus, forces on distant MM point charges are essentially correct. If, however, the point charges penetrate into the QM density, they are partially shielded, and their gradients (forces) differ from those obtained from the ESP charges. In Figure 7, the differences between the absolute values of the electric fields caused by the ESP charges, $|\vec{E}_{ESP}|$, and the density, $|\vec{E}_{\rho}|$, at the position of the MM atoms of the window with $\xi_i = -0.4$ Å are shown. The electrostatic force on the MM atom i with charge Q_i is obtained as $\vec{F}_i = Q_i\vec{E}$.

As a more spatially extended density is expected to lead to a more pronounced shielding, we calculated the effect not only for AM1 but also for BP86/aug-cc-pVTZ, see Figure 7. The values are obtained from geometries with $\xi_i = -0.4$ Å optimized with AM1 and BP86/TZVP, respectively. It can be seen from Figure 7 that the extended basis set, which includes diffuse functions, leads only to slightly larger errors in the electrostatic forces than the minimum-basis set of AM1. Thus, we expect that ESP charges are also a good approximation for extended basis sets.

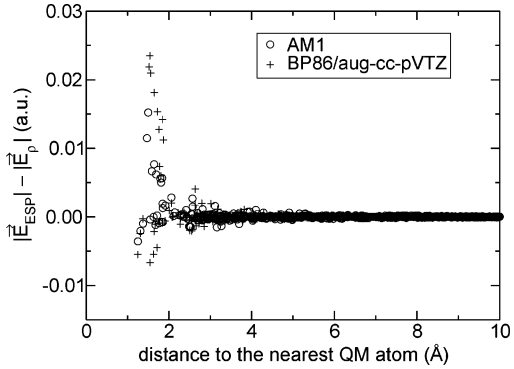


Figure 7. Differences between the absolute values of the electric fields caused by the ESP charges, $|\vec{E}_{ESP}|$, and the density, $|\vec{E}_{\rho}|$. The difference is evaluated at the positions of the MM atoms.

QM/MM-FEP with full SCF iterations considers all energy contributions that enter the QTCP-U approach of Rod and Ryde.[22] Our finding that ESP charges are a good approximation is in agreement with their work.

III.C.3. Link Atoms. The method used to treat the link atoms does not influence the results appreciably, as seen from Table 4 (last four lines). There is only one junction between the QM part and the MM part in our system; thus, PHBH may not represent a severe test for the link atom treatment. However, its influence on the free energy is so small that we do not expect a significant effect even if the system contains more link atoms. The simulations have been done with the full frozen density. The differences between the simulations are caused only by the link atom treatment as we sampled along the same trajectories. This was achieved by starting from the same configuration and velocity distribution. The link atom treatment only affects the perturbation, but not the dynamics. A comparison of the geometry after the simulation verified that the same trajectories had been sampled. Because of the small effect of the link atom treatment, see Table 4, formal aspects (section II.D) recommend method 4.

IV. Conclusion

We used the example of PHBH to show that QM/MM-FEP is a reliable and efficient method to calculate reaction free energies and free-energy barriers. We tested several methodological approximations on the system PHBH. Freezing the density during the FEP sampling caused an error of 2.5 kJ mol^{-1} in the computed barrier. This error was even smaller, 1.9 kJ mol^{-1}, when the frozen density was approximated by point charges. Different choices of the link-atom treatment altered the barrier by up to 1.5 kJ mol^{-1}. These numbers should be judged in the light of a typical sampling error of 1.0 kJ mol^{-1} in our simulations. We find that it is advisable to perturb all three atoms of a QM/MM junction (method 4) and that it is adequate to approximate the QM density by ESP charges in the FEP sampling.

When ESP charges are used, the computer time required for the geometry optimizations exceeds the time required for the MD simulations when using a demanding QM method. Thus, QM/MM-FEP is affordable at any level of QM theory where one can afford the geometry optimizations. The choice of the reaction coordinates is only limited by the optimizer. For the HDLC optimizer used in this study, it is straightforward to implement any linear combination of internal coordinates as a constraint. Intrinsically, constraints are unnecessary in the MD simulation, with the exception of freezing the Cartesian coordinates of the QM part and the first MM atom of each junction. This junction atom should then be included in the calculation of the Hessian for the harmonic approximation of the QM entropy.

Acknowledgment. This work was supported by the Volkswagenstiftung, Grant I/80454, and the Deutsche Forschungsgemeinschaft, Grant SFB 663/C4.

References

(1) Torrie, G. M.; Valleau, J. P. *Chem. Phys. Lett.* **1974**, *28*, 578.

(2) Torrie, G. M.; Valleau, J. P. *J. Comput. Phys.* **1977**, *23*, 187.

(3) Kästner, J.; Thiel, W. *J. Chem. Phys.* **2005**, *123*, 144104.

(4) den Otter, W. K.; Briels, W. J. *J. Chem. Phys.* **1998**, *109*, 4139.

(5) Kirkwood, J. G. *J. Chem. Phys.* **1935**, *3*, 300.

(6) Schlitter, J.; Klähn, M. *J. Chem. Phys.* **2003**, *118*, 2057.

(7) Sprik, M.; Ciccotti, G. *J. Chem. Phys.* **1998**, *109*, 7737.

(8) Zhang, Y.; Liu, H.; Yang, W. *J. Chem. Phys.* **2000**, *112*, 3483.

(9) Bentzien, J.; Muller, R. P.; Florián, J.; Warshel, A. *J. Phys. Chem. B* **1998**, *102*, 2293.

(10) Strajbl, M.; Hong, G.; Warshel, A. *J. Phys. Chem. B* **2002**, *106*, 13333.

(11) Chandrasekhar, J.; Smith, S. F.; Jorgensen, W. L. *J. Am. Chem. Soc.* **1984**, *106*, 3049.

(12) Chandrasekhar, J.; Smith, S. F.; Jorgensen, W. L. *J. Am. Chem. Soc.* **1985**, *107*, 154.

(13) Jorgensen, W. L. *Acc. Chem. Res.* **1989**, *22*, 184.

(14) Stanton, R. V.; Perakyla, M.; Bakowies, D.; Kollman, P. A. *J. Am. Chem. Soc.* **1998**, *120*, 3448.

(15) Kuhn, B.; Kollman, P. A. *J. Am. Chem. Soc.* **2000**, *122*, 2586.

(16) Donini, O.; Darden, T.; Kollman, P. A. *J. Am. Chem. Soc.* **2000**, *122*, 12270.

(17) Kollman, P. A.; Kuhn, B.; Donini, O.; Perakyla, M.; Stanton, R.; Bakowies, D. *Acc. Chem. Res.* **2001**, *34*, 72.

(18) Zwanzig, R. W. *J. Chem. Phys.* **1954**, *22*, 1420.

(19) Cisneros, G. A.; Liu, H.; Zhang, Y.; Yang, W. *J. Am. Chem. Soc.* **2003**, *125*, 10384.

(20) Liu, H.; Zhang, Y.; Yang, W. *J. Am. Chem. Soc.* **2000**, *122*, 6560.

(21) Rod, T. H.; Ryde, U. *Phys. Rev. Lett.* **2005**, *94*, 138302.

(22) Rod, T. H.; Ryde, U. *J. Chem. Theory Comput.* **2005**, *1*, 1240.

(23) Entsch, B.; van Berkel, W. J. *FASEB J.* **1995**, *9*, 476.

(24) Jadan, A. P.; Moonen, M. J. H.; Boeren, S.; Golovleva, L. A.; Rietjens, I. M. C. M.; van Berkel, W. J. H. *Adv. Synth. Catal.* **2004**, *346*, 367.

(25) Entsch, B.; Ballou, D. P.; Massey, V. *J. Biol. Chem.* **1976**, *251*, 2550.

(26) Entsch, B.; Palfey, B. A.; Ballou, D. P.; Massey, V. *J. Biol. Chem.* **1991**, *266*, 17341.

(27) Husain, M.; Entsch, B.; Ballou, D. P.; Massey, V.; Chapman, P. J. *J. Biol. Chem.* **1980**, *255*, 4189.

(28) van Berkel, W. J. H.; Müller, F. *Eur. J. Biochem.* **1989**, *179*, 307.

(29) Billeter, S. R.; Hanser, C. F. W.; Mordasini, T. Z.; Scholten, M.; Thiel, W.; van Gunsteren, W. F. *Phys. Chem. Chem. Phys.* **2001**, *3*, 688.

(30) Ridder, L.; Mulholland, A. J.; Vervoort, J.; Rietjens, I. M. C. M. *J. Am. Chem. Soc.* **1998**, *120*, 7641.

(31) Ridder, L.; Mulholland, A. J.; Rietjens, I. M. C. M.; Vervoort, J. *J. Mol. Graphics Modell.* **1999**, *17*, 163.

(32) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. *J. Am. Chem. Soc.* **1985**, *107*, 3902.

(33) Atkins, P. W.; de Paula, J. *Physical Chemistry*, 7th ed.; Oxford University Press: Oxford, U. K., 2002; Chapter 19.

(34) Sherwood, P.; de Vries, A. H.; Guest, M. F.; Schreckenbach, G.; Catlow, C. R. A.; French, S. A.; Sokol, A. A.; Bromley, S. T.; Thiel, W.; Turner, A. J.; Billeter, S.; Terstegen, F.; Thiel, S.; Kendrick, J.; Rogers, S. C.; Casci, J.; Watson, M.; King, F.; Karlsen, E.; Sjøvoll, M.; Fahmi, A.; Schäfer, A.; Lennartz, C. *THEOCHEM* **2003**, *632*, 1.

(35) Senn, H. M.; Thiel, S.; Thiel, W. *J. Chem. Theory Comput.* **2005**, *1*, 494.

(36) Gatti, D. L.; Entsch, B.; Ballou, D. P.; Ludwig, M. L. *Biochemistry* **1996**, *35*, 567.

(37) Thiel, W. *MNDO99*, version 6.1; Max-Planck-Institut für Kohlenforschung: Mülheim an der Ruhr, Germany, 2004.

(38) van Gunsteren, W. F.; Billeter, S. R.; Eising, A. A.; Hünenberger, P. H.; Krüger, P.; Mark, A. E.; Scott, W. R. P.; Tironi, I. G. *Biomolecular Simulation: The GROMOS96 Manual and User Guide*; vdf and BIOMOS b.v.: Zürich, Switzerland; Groningen, The Netherlands, 1996.

(39) Tironi, I. G.; Sperb, R.; Smith, P. E.; van Gunsteren, W. F. *J. Chem. Phys.* **1995**, *102*, 5451.

(40) Baker, J.; Kessi, A.; Delley, B. *J. Chem. Phys.* **1996**, *105*, 192.

(41) Billeter, S. R.; Turner, A. J.; Thiel, W. *Phys. Chem. Chem. Phys.* **2000**, *2*, 2177.

(42) Liu, D. C.; Nocedal, J. *Math. Prog. B* **1989**, *45*, 503.

(43) Banerjee, A.; Adams, N.; Simons, J.; Shepard, R. *J. Phys. Chem.* **1985**, *89*, 52.

(44) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684.

(45) Hoover, W. G. *Phys. Rev. A* **1985**, *31*, 1695.

(46) Martyna, G. J.; Klein, M. L.; Tuckerman, M. *J. Chem. Phys.* **1992**, *97*, 2635.

(47) Nosé, S. *Mol. Phys.* **1984**, *52*, 255.

(48) Nosé, S. *J. Chem. Phys.* **1984**, *81*, 511.

(49) Jang, S.; Voth, G. A. *J. Chem. Phys.* **1997**, *107*, 9514.

(50) Ryckaert, J.-P.; Ciccottiand, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *23*, 327.

(51) Schiferl, S. K.; Wallace, D. C. *J. Chem. Phys.* **1985**, *83*, 5203.

(52) Mann, H. B. *Econometrica* **1945**, *13*, 245.

(53) Kumar, S.; Rosenberg, J. M.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A. *J. Comput. Chem.* **1992**, *13*, 1011.

(54) Souaille, M.; Roux, B. *Comput. Phys. Commun.* **2001**, *135*, 40.

(55) Hohenberg, P.; Kohn, W. *Phys. Rev.* **1964**, *136*, B864.

(56) Kohn, W.; Sham, L. J. *Phys. Rev.* **1965**, *140*, A1133.

(57) Ahlrichs, R.; Bär, M.; Häser, M.; Horn, H.; Kölmel, C. *Chem. Phys. Lett.* **1989**, *162*, 165.

(58) Eichkorn, K.; Treutler, O.; Öhm, H.; Häser, M.; Ahlrichs, R. *Chem. Phys. Lett.* **1995**, *242*, 652.

(59) Eichkorn, K.; Weigend, F.; Treutler, O.; Ahlrichs, R. *Theor. Chem. Acc.* **1997**, *97*, 119.

(60) Sierka, M.; Hogekamp, A.; Ahlrichs, R. *J. Chem. Phys.* **2003**, *118*, 9136.

(61) Treutler, O.; Ahlrichs, R. *J. Chem. Phys.* **1995**, *102*, 346.

(62) Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098.

(63) Dirac, P. A. M. *Proc. R. Soc. London, Ser. A* **1929**, *123*, 714.

(64) Perdew, J. P. *Phys. Rev. B* **1986**, *33*, 8822.

(65) Slater, J. C. *Phys. Rev.* **1951**, *81*, 385.

(66) Vosko, S. H.; Wilk, L.; Nusair, M. *Can. J. Phys.* **1980**, *58*, 1200.

(67) Schäfer, A.; Huber, C.; Ahlrichs, R. *J. Chem. Phys.* **1994**, *100*, 5829.

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

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

(70) Jarzynski, C. *Phys. Rev. Lett.* **1997**, *78*, 2690.

(71) Kenney, J. F.; Keeping, E. S. *Mathematics of Statistics*; Van Nostrand: Princeton, NJ, 1951; Vol. 2.

(72) Marcinkiewicz, J. *Math. Z.* **1939**, *44*, 612.

(73) Park, S.; Khalili-Araghi, F.; Tajkhorshid, E.; Schulten, K. *J. Chem. Phys.* **2003**, *119*, 3559.

(74) Weisstein, E. W. k-Statistic; MathWorld-A Wolfram Web Resource. http://mathworld.wolfram.com/k-Statistic.html (accessed Aug 26, 2005).

CT050252W